PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 19/00 — A1

(11) International Publication Number: **WO 97/27559**

(43) International Publication Date: 31 July 1997 (31.07.97)

(21) International Application Number: PCT/US97/01491

(22) International Filing Date: 27 January 1997 (27.01.97)

(30) Priority Data:
08/592,132 — 26 January 1996 (26.01.96) — US
08/657,147 — 3 June 1996 (03.06.96) — US

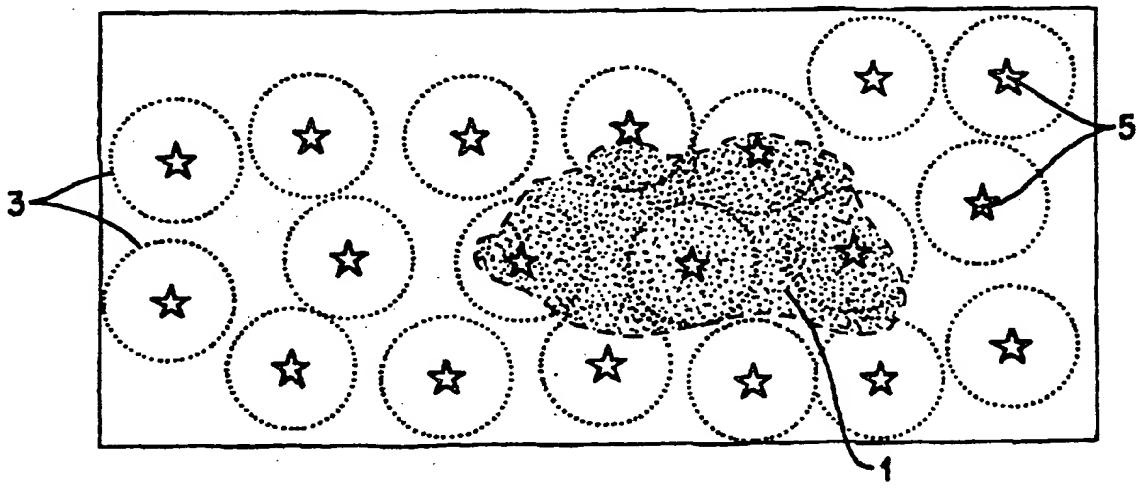
(71)(72) Applicants and Inventors: PATTERSON, David, E. [US/US]; 1908 Bookbinder Drive, St. Louis, MO 63146 (US). CRAMER, Richard, D. [US/US]; 9012 Highway DD, O'Fallon, MO 63366 (US). CLARK, Robert, D. [US/US]; 827 Renee Lane, St. Louis, MO 63141 (US). FERGUSON, Allan, M. [GB/US]; 2314 Callender Court, St. Louis, MO 63017 (US).

(74) Agent: WEINBERGER, Laurence, A.; Suite 103, 882 S. Matlack Street, West Chester, PA 19382 (US).

(81) Designated States: AU, CA, CN, CZ, HU, IL, JP, KR, NO, PL, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD OF CREATING AND SEARCHING A MOLECULAR VIRTUAL LIBRARY USING VALIDATED MOLECULAR STRUCTURE DESCRIPTORS



(57) Abstract

The problem of how to select out of a large chemically accessible universe molecules representative of the diversity of that universe is resolved by the discovery of a method to validate m lecular structural descriptors. Using the validated descriptors, optimally diverse subsets (5) can be selected. In addition, from the universe, molecules with characteristics similar to a selected molecule can be identified (3). The validated descriptors also enable the generati n of a huge virtual library of potential product molecules which could be formed by combinatorial arrangement of structural variations and cores. In this virtual library it is possible to search billions of possible product compounds in relatively short time frames.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	**GB**	United Kingdom	**MW**	Malawi
AT	Austria	**GE**	Georgia	**MX**	Mexico
AU	Australia	**GN**	Guinea	**NE**	Niger
BB	Barbados	**GR**	Greece	**NL**	Netherlands
BE	Belgium	**HU**	Hungary	**NO**	Norway
BF	Burkina Faso	**IE**	Ireland	**NZ**	New Zealand
BG	Bulgaria	**IT**	Italy	**PL**	Poland
BJ	Benin	**JP**	Japan	**PT**	Portugal
BR	Brazil	**KE**	Kenya	**RO**	Romania
BY	Belarus	**KG**	Kyrgystan	**RU**	Russian Federation
CA	Canada	**KP**	Democratic People's Republic of Korea	**SD**	Sudan
CF	Central African Republic			**SE**	Sweden
CG	Congo	**KR**	Republic of Korea	**SG**	Singapore
CH	Switzerland	**KZ**	Kazakhstan	**SI**	Slovenia
CI	Côte d'Ivoire	**LI**	Liechtenstein	**SK**	Slovakia
CM	Cameroon	**LK**	Sri Lanka	**SN**	Senegal
CN	China	**LR**	Liberia	**SZ**	Swaziland
CS	Czechoslovakia	**LT**	Lithuania	**TD**	Chad
CZ	Czech Republic	**LU**	Luxembourg	**TG**	Togo
DE	Germany	**LV**	Latvia	**TJ**	Tajikistan
DK	Denmark	**MC**	Monaco	**TT**	Trinidad and Tobago
EE	Estonia	**MD**	Republic of Moldova	**UA**	Ukraine
ES	Spain	**MG**	Madagascar	**UG**	Uganda
FI	Finland	**ML**	Mali	**US**	United States of America
FR	France	**MN**	Mongolia	**UZ**	Uzbekistan
GA	Gabon	**MR**	Mauritania	**VN**	Viet Nam

METHOD OF CREATING AND SEARCHING A MOLECULAR VIRTUAL LIBRARY USING VALIDATED MOLECULAR STRUCTURE DESCRIPTORS

Technical Field

This invention relates to the field of molecular structure/activity analysis and more specifically to: 1) a method of validating molecular structural descriptors; 2) a method using validated molecular descriptors to design an optimally diverse combinatorial screening library; 3) a method of merging libraries derived from different combinatorial chemistries; 4) a method using validated molecular descriptors of generating a searchable virtual library of molecules which can be combinatorially derived; 5) methods of searching the virtual library for combinatorially derived product molecules which meet specified criteria; and 6) methods of following up and optimizing identified leads. The screening libraries designed by the methods of this invention are constructed to ensure that an optimal structural diversity of compounds is represented. The search methods of the invention ensure that the same diversity space is not oversampled and that compounds can be identified having a high likelihood of possessing the same structure and/or activity of a lead compound. In particular, the invention describes the design of libraries of small molecules to be used for pharmacological testing.

Background Art

Statement Of The Problem

While the present invention is discussed with detailed reference to the search for and identification of pharmacologically useful chemical compounds, the invention is applicable to any attempt to search for and identify chemical compounds which have some desired physical or chemical characteristic(s). The broader teachings of this invention are easily recognized if a different functional utility or useful property describing other chemical systems is substituted below for the term "biological activity".

Starting with the serendipitous discovery of penicillin by Fleming and the subsequent directed searches for additional antibiotics by Waksman and Dubos, the field of drug discovery during the post World War II era has been driven by the belief that nature would provide many needed drugs if only a careful and diligent search for them was conducted. Consequently, pharmaceutical companies undertook massive screening programs which tested samples of natural products (typically isolated from soil or plants) for their biological properties. In a parallel effort to increase the effectiveness of the discovered "lead" compounds, medicinal chemists learned to synthesize derivatives and analogs of the compounds. Over the years, as biochemists identified new enzymes and biological reactions, large scale screening continued as compounds were tested for biological activity in an ever rapidly expanding number of biochemical pathways. However, proportionately fewer and fewer lead compounds possessing a desired therapeutic activity have been discovered. In an attempt to extend the range of compounds available for testing, during the last few years the search for unique biological materials has been extended to all corners of the earth including sources from both the tropical rain forests and the ocean. Despite these and other efforts, it is estimated that discovery and development of each new drug still takes about 12 years and costs on the order of 350 million dollars.

Beginning approximately twenty-five years ago, as bioscientists learned more about the chemical and stereochemical requirements for biological interactions, a variety of semi-empirical, theoretical, and quantitative approaches to drug design were developed. These approaches were accelerated by the availability of powerful computers to perform computational chemistry. It was hoped that the era of "rational drug design" would shorten the time between significant discoveries and also provide an approach to discovering compounds active in biological pathways for which no drugs had yet been discovered. In large part, this work was based on the accumulated observation of medicinal chemists that compounds which were structurally similar also possessed similar biological activities. While significant strides were made using this approach, it too, like the mass screening programs, failed to provide a solution to the problem of rapidly discovering new compounds with activities in the ever increasing number of biological pathways being elucidated by modern biotechnology.

During the past four or five years, a revised screening approach has been under development which, it was hoped, would accelerate the pace of drug discovery. In fact, the approach has been remarkably successful and represents one of the most active areas in biotechnology today. This new approach utilizes combinatorial libraries against which

biological assays are screened. Combinatorial libraries are collections of molecules generated by synthetic pathways in which either: 1) two groups of reactants are combined to form products; or 2) one or more positions on core molecules are substituted by a different chemical constituent/moiety selected from a large number of possible constituents.

Two fundamental ideas underlie combinatorial screening libraries. The first idea, common to all drug research, is that somewhere amongst the diversity of all possible chemical structures there exist molecules which have the appropriate shape and binding properties to interact with any biological system. The second idea is the belief that synthesizing and testing many molecules in parallel is a more efficient way (in terms of time and cost) to find a molecule possessing a desired activity than the random testing of compounds, no matter what their source. In the broadest context, these ideas require that, since the binding requirements of a ligand to the biological systems under study (enzymes, membranes, receptors, antibodies, whole cell preparations, genetic materials, etc.) are not known, the screened compounds should possess as broad a range of characteristics (chemical and physical) as possible in order to increase the likelihood of finding one that is appropriate for any given biological target. This requirement for a screening library is reflected in the term "diversity" - essentially a way of suggesting that the library should contain as great a dissimilarity of compounds as possible.

However, as is immediately apparent, a combinatorial approach to synthesizing molecules generates an immense number of compounds many with a high degree of structural similarity. In fact, the number of compounds synthetically accessible with known organic reactions exceeds by many orders of magnitude the numbers which can actually be made and tested. One area where these ideas were first explored is in the design of peptide libraries. For a library of five member peptides synthesized using the 20 naturally occurring amino acids, 3,200,000, (20^5) different peptides may be constructed. The number of combinatorial possibilities increases even more dramatically when non-peptide combinatorial libraries are considered. With non-peptide libraries, the whole synthetic chemical universe of combinatorial possibilities is available. Library sizes ranging from 5×10^7 to 4×10^{12} molecules are now being discussed. The enormous universe of chemical compounds is both a blessing and a curse to medicinal chemists seeking new drugs. On the one hand, if a molecule exists with the desired biological activity, it should be included in the chemical universe. On the other hand, it may be impossible to find. Thus, the principal focus of recent efforts has been to define smaller screening subsets of molecules derivable from accessible combinatorial syntheses without losing the inherent diversity of an accessible universe.

To date, in order to narrow the focus of the search and reduce the number of compounds to be screened, attention has been directed to designing biologically specific libraries. Thus, many combinatorial screening libraries existing in the prior art have been designed based on prior knowledge about a particular biological system such as a known pharmacophore (a geometric arrangement of structural fragments abstracted from molecular structures known to have activity). Even with this knowledge, molecules are included in these prior art libraries based on intuition - "seat of the pants" estimations of likely similarity based on an intuitive "feel" for the systems under study. This procedure is essentially pseudo-random screening, not rational library design. Several biotechnology startup companies have developed just such proprietary libraries, and success using combinatorial libraries has been achieved by sheer effort. In one example 18 libraries containing 43 million compounds were screened to identify 27 active compounds[1]. With library searches of this magnitude, it is most likely that the enormous number of inactive molecules [(43 X 10^6) - 27] must have included staggering numbers of redundantly inactive molecules - molecules not significantly distinguishable from one another - even in libraries designed with a particular biological target in mind. Clearly, when searching for a lead molecule which interacts with an uncharacterized biological target, approaches requiring knowledge of the biological targets will not work. But finding such a lead is exactly the case for which it is hoped general purpose screening libraries can be designed. If the promise of combinatorial chemistry is ever to be fully realized, some rational and quantitative method of reducing the astronomical number of compounds accessible in the combinatorial chemistry universe to a number which can be usefully tested is required. In other words, the efficiency of the search process must be increased. For this purpose, a smaller rationally designed screening library, which still retains the diversity of the combinatorially accessible compounds, is absolutely necessary.

Thus, there are two criteria which must be met by any screening library subset of some universe of combinatorially accessible compounds. First, the diversity, the dissimilarity of the universe of compounds accessible by some combinatorial reaction, must be retained in the screening subset. A subset which does not contain examples of the total range of diversity in such a universe would potentially miss critical molecules, thereby frustrating the very reason for the creation of the subset. Second, for efficient screening, the ideal subset should not contain more than one compound representative of each aspect of the diversity of the larger group. If more than one example were included, the same diversity would be tested more than once. Such redundant screening would yield no new information while simultaneously

increasing the number of compounds which must be synthesized and screened. Therefore, the fundamental problem is how to reduce to a manageable number the number of compounds that need to be synthesized and tested while at the same time providing a reasonably high probability that no possible molecule of biological importance is overlooked. (In this regard, it should be recognized that the only way of absolutely insuring that all diversity is represented in a library is to include and test all compounds.) A conceptual analogy to the problem might be: what kind of filter can be constructed to sort out from the middle of a blinding snowstorm individual snowflakes which represent all the classes of crystal structures which snowflakes can form?

The fundamental question plaguing progress in this area has been whether the concept of the diversity of molecular structure can be usefully described and quantified; that is, how is it possible to compare/distinguish the physical and chemical properties determinative of biological activity of one molecule with that of another molecule? Without some way to quantitatively describe diversity, no meaningful filter can be constructed. Fortunately, for biological systems, the accumulated wisdom of bioscientists has recognized a general principle alluded to earlier which provides a handle on this problem. As framed by Johnson and Maggiora[2], the principle is simply stated as: "structurally similar molecules are expected to exhibit similar (biological) properties." Based on this principle, quantifying diversity becomes a matter of quantifying the notion of structural similarity. Thus, for design of a screening subset of a combinatorial library (hereafter referred to as a "combinatorial screening library"), it should only be necessary to identify which molecules are structurally similar and which structurally dissimilar. According to the selection criteria outlined above, one molecule of each structurally similar group in the combinatorially accessible chemical universe would be included in the library subset. Such a library would be an optimally diverse combinatorial screening library. The problem for medicinal chemists is to determine how the intuitively perceived notions of structural similarity of chemical compounds can be validly quantified. Once this question is satisfactorily answered, it should be possible to rationally design combinatorial screening libraries.

Prior Art Approaches

Many descriptors of molecular structure have been created in the prior art in an attempt to quantify structural similarity and/or dissimilarity. As the art has recognized, however, no method currently exists to distinguish those descriptors that quantify useful aspects of similarity from those which do not. The importance of being able to validate molecular descriptors has

been a vexing problem restricting advances in the art, and, before this invention, no generally applicable and satisfactory answer had been found. The problem may be conceptualized in terms of a multidimensional space of structurally derivable properties which is populated by all possible combinatorially accessible chemical compounds. Compounds lying "near" one another in any one dimension may lie "far apart" from one another in another dimension. The difficulty is to find a useful design space - a quantifiable dimensional space (metric space) in which compounds with similar biological properties cluster; ie., are found measurably near to each other. What is desired is a molecular structural descriptor which, when applied to the molecules of the chemical universe, defines a dimensional space in which the "nearness" of the molecules with respect to a specified characteristic (ie.; biological activity) in the chemical universe is preserved in the dimensional space. A molecular structural descriptor (metric) which does not have this property is useless as a descriptor of molecular diversity. A valid descriptor is defined as one which has this property.

In light of the above, it should be noted that there is a difference between a descriptor being valid and being perfect. There may or may not be a "perfect" metric which precisely and quantitatively maps the diversity of compounds (much less those of biological interest). However, a good approximation is sufficient for purposes of designing a combinatorial screening library and is considered valid/useful. Acceptance of this validation/usefulness criteria is essentially equivalent to saying that, if there is a high probability that if one molecule is active (or inactive), a second molecule is also active (or inactive), then most of the time sampling one of the pair will be sufficient. Restating this same principle with a slightly different emphasis highlights another feature, namely: the design criteria for combinatorial screening libraries should yield a high probability that, for any given inactive molecule, it is more probable to find an active molecule somewhere else rather than as a near neighbor of that inactive molecule. While this is a probabilistic approach, it emphasizes that a good approximation to a perfect metric is sufficient for purposes of designing a combinatorial screening library as well as in other situations where the ability to discriminate molecular structural difference and similarities is required. A perfect descriptor (certainty) for pharmacological searching is not needed to achieve the required level of confidence as long as it is valid (maps a subspace where biological properties cluster).

The typical prior art approach for establishing selection criteria for screening library subsets relied on the following clustering paradigm: 1) characterization of compounds according to a chosen descriptor(s) (metric[s]); 2) calculation of similarities or "distances" in

the descriptor (metric) between all pairs of compounds; and 3) grouping or clustering of the compounds based on the descriptor distances. The idea behind the paradigm is that, within a cluster, compounds should have similar activities and, therefore, only one or a few compounds from each cluster, which will be representative of that cluster, need be included in a library. The actual clustering is done until the prior art user feels comfortable with the groupings and their spacing. However, with no knowledge of the validity/usefulness of the descriptor employed, and no guidance with respect to the size or spacing of clusters to be expected from any given descriptor, prior art clustering has been, at best, another intuitive "seat of the pants" approach to diversity measurement.

The prior art describes the construction and application of many molecular structural descriptors while all the while tacitly acknowledging that little progress has been made towards solving the fundamental problem of establishing their validity. The field has nevertheless proceeded based on the belief/faith that, by incorporating in the descriptors certain measures which had been recognized in QSAR studies as being important contributors to defining structure-activity relationships, valid/useful descriptors would be produced. In a leading method representative of this prior art approach to defining a similarity descriptor, E. Martin et al.[3] construct a metric for quantifying structural similarity using measures that characterize lipophilicity, shape and branching, chemical functionality, and receptor recognition features. (For the reasons set forth later in relation to the present invention, Martin et al. applied their metric to the reactants which would be used in combinatorial synthesis.) This large set of measures is used to generate a statistically blended metric consisting of a total of 16 properties for each individual reactant studied (5 shape descriptors, 5 measures of chemical functionality, 5 receptor binding descriptors, and one lipophilicity property). This generates a 16 dimensional property space. The 16 properties are simultaneously displayed in a circular "Flower Plots" graphical environment, where each property is assigned a petal. All the plots together visually display how the diversity of the studied reactants is distributed through the computed property space. Martin acknowledges that the plots "...cannot, of course, prove that the subset is diverse in any 'absolute' sense, independent of the calculated properties." (Martin at 1434)

In another approach relating to peptoid design, Martin et al.[4] have characterized the varieties of shape that an unknown receptor cavity might assume by a few assemblages of blocks, called "polyominos". Candidates for a combinatorial design are classified by the types of polyominos into which they can be made to fit, or "docked". The 7 flexible polyomino shape descriptors are added to the previously defined 16 descriptors to yield a 23 dimensional

property space. Martin has demonstrated that the docking procedure generates for a methotrexate ligand in a cavity of dihydrofolate reductase nearly the correct structure as that established by X-ray diffraction studies. The docking procedure, which must be applied to every design candidate for each polyomino, requires a considerable amount of CPU time (is computationally expensive). However, a problem with this approach is the conceptually severe (unjustified) approximation of representing all possible irregularly shaped receptor cavities by only about a dozen assemblies of smooth-sided polyomino cubes. Martin has also presented no validation of the approach, which in this case, would be a demonstration that molecules which fit into the same polyominos tend to have similar biological properties.

One approach which has been taken to try to empirically assess the relative validity of prior art metrics has been to survey the metrics to see if any of them appeared to be superior to any others as judged by clustering analysis. Y. C. Martin et al.[5] have reported that 3D fingerprints, collections of fragments defined by pairs of atoms and their accessible interatomic distances, perform no better than collections of 2D fragments in defining clusters that separate biologically active from inactive compounds. As will be seen later, some of this work pointed towards the possible validity of one metric, but the authors concentrated on the comparative clustering aspects and did not follow up on the broader import of the data.

W. Herndon[6] among others has pointed out that an experimentally determined similarity QSAR is, by definition, a good test of the validity of that similarity concept for the biological system from which it is derived and may have some usefulness in estimating diversity for that system. However, QSARs essentially map only the space of a particular receptor, do not provide information about the validity of other descriptors, and would be generally inapplicable to construction of a combinatorial screening library designed for screening unknown receptors or those for which no QSAR data was available.

Finally, D. Chapman et al.[7] have used their "Compass" 3D-QSAR descriptor which is based on the three dimensional shape of molecules, the locations of polar functionalities on the molecules, and the fixation entropies of the molecules to estimate the similarity of molecules. Essentially, using the descriptor, they try to find the molecules which have the maximum overlap (in geometric/cartesian space) with each other. The shape of each molecule of a series is allowed to translate and rotate relative to each other molecule and the internal degrees of freedom are also allowed to rotate in an iterative procedure until the shapes with greatest or least overlap similarity are identified. Selecting 20 maximally diverse carboxylic acids based on seeking the maximally diverse alignment of each of the 3000 acids considered took

approximately 4 CPU computing weeks by their method. No indication was given of whether their descriptor was valid in the sense defined above, and, clearly, such a procedure would be too time consuming to apply to a truly large combinatorial library design.

One way in which many of the prior art approaches attempt to work around the problem of not knowing if a molecular structural descriptor is valid is to try, when clustering, to maximize as much as possible the distance between the clusters from which compounds will be selected for inclusion in the screening library subset. The thinking behind this approach is that, if the clusters are far enough apart, only molecules diverse from each other will be chosen. Conversely, it is thought that, if the clusters are close together, oversampling (selection of two or more molecules representative of the same elements of diversity) would likely occur. However, as we have seen, if the metric used in the cluster analysis is not initially valid (does not define a subspace in which molecules with similar biological activity cluster), then no amount of manipulation will prevent the sample from being essentially random. Worse yet, an invalid metric might not yield a selection as good as random! The acknowledgement by Martin quoted above is a recognition of the prior art's failure to yet discover a general method for validating descriptors.

Another related problem in the prior art is the failure to have any objective manner of ascertaining when the library subset under design has an adequate number of members; that is, when to stop sampling. Clearly, if nothing is known about the distribution of the diversity of molecules, one arbitrary stopping point is as good as any other. Any stopping point may or may not sample sufficiently or may oversample. In fact, the prior art has not recognized a coherent quantitative methodology for determining the end point of selection. Essentially, in the prior art, a metric is used to maximize the presumed differences between molecules (typically in a clustering analysis), and a very large number of molecules are chosen for inclusion in a screening library subset based on the belief that there is safety in numbers; that sampling more molecules will result in sampling more of the diversity of a combinatorially accessible chemical space. As pointed out earlier, however, only by including all possible molecules in a library will one guarantee that all of the diversity has been sampled. Short of such total sampling, users of prior art library subsets constructed along the lines noted above do not know whether a random sample, a representative sample, or a highly skewed sample has been screened.

Several other problems flow from the inability to rationally select a combinatorial screening library for optimal diversity and these are related both to the chemistry used to

create the combinatorial library and the screening systems used. First, because many more molecules may have to be synthesized than may be needed, mass synthetic schemes have to be devised which create many combinations simultaneously. In fact, there is a good deal of disagreement in the prior art as to whether compounds should be synthesized individually or collectively or in solution or on solid supports. Within any synthetic scheme, an additional problem is keeping track of and identifying the combinations created. It should be understood that, where relatively small (molecular weight of less than about 1500) organic molecules are concerned, generally standard, well known, organic reactions are used to create the molecules. In the case of peptide like molecules, standard methods of peptide synthesis are employed. Similarly for polysaccharides and other polymers, reaction schemes exist in the prior art which are well known and can be utilized. While the synthesis of any individual combinatorial molecule may be straightforward, much time and effort has been and is still being expended to develop synthetic schemes in which hundreds, thousands, or tens of thousands of combinatorial combinations can be synthesized simultaneously.

In many synthetic schemes, mixtures of combinatorial products are synthesized for screening in which the identity of each individual component is uncertain. Alternatively, many different combinatorial products may be mixed together for simultaneous screening. Each additional molecule added to a simultaneous screen means that many fewer individual screening operations have to be performed. Thus, it is not unusual that a single assay may be simultaneously tested against up to 625 or more different molecules. Not until the mixture shows some activity in the biological screening assay will an attempt be made to identify the components. Many approaches in the prior art therefore face "deconvolution" problems; ie. trying to figure out what was in an active mixture either by following the synthetic reaction pathway, by resynthesizing the individual molecules which should have resulted from the reaction pathway, or by direct analysis of duplicate samples. Some approaches even tag the carrier of each different molecule with a unique molecular identifier which can be read when necessary. All these problems are significantly decreased by designing a library for optimal diversity.

Another major problem with the inclusion of multiple and potentially non-diverse compounds in the same screening mixture is that many assays will yield false positives (have an activity detected above a certain established threshold) due to the combined effect of all the molecules in the screening mixture. The absence of the desired activity is only determined after expending the time, effort, and expense of identifying the molecules present in the mixture and

testing them individually. Such instances of combined reactivity are reduced when the screening mixture can be selected from molecules belonging to diverse groups of an optimally designed library since it is not as likely that molecules of different (diversity) structures would likely produce a combined effect.

It is clear that a great deal of cleverness has been expended in actually manufacturing the combinatorial libraries. While the basic chemistry of synthesizing any given molecule is straight forward, the next advance in the development of combinatorial chemistry screening libraries will be optimization of the design of the libraries.

Further problems in the prior art arise in the attempt to follow up leads resulting from the screening process. As noted above, many libraries are designed with some knowledge of the receptor and its binding requirements. While, within those constraints, all possible combinatorial molecules are synthesized for screening, finding a few molecules with the desired activity among such a library yields no information about what active molecules might exist in the universe accessible with the same combinatorial chemistry but outside the limited (receptor) library definition. This is an especially troubling problem since, from serendipitous experience, it is well known that sometimes totally unexpected molecules with little or no obvious similarity to known active molecules exhibit significant activity in some biological systems. Thus, even finding a candidate lead in a library whose design was based on knowledge of the receptor is no guarantee that the lead can be followed to an optimal compound. Only a rationally designed combinatorial screening library of optimal diversity can approach this goal.

For prior art library subsets designed around the use of some descriptor to cluster compounds, similar problems may exist. In such a library design, one or at most a few compounds will have been selected from each cluster. Only if the descriptor is valid, does such a selection procedure make sense. If the descriptor is not valid, each cluster will contain molecules representative of many different diversities and selecting from each cluster will still have resulted in a random set of molecules which do not sample all of the diversity present. Since the prior art does not possess a generally applicable method of validating descriptors, all screening performed with prior art libraries is suspect and may not have yielded all the useful information desired about the larger chemical universe from which the library subsets were selected.

Finally, as the expense in time and effort of creating and screening combinatorial libraries increases, the question of the uniqueness of the libraries becomes ever more critical.

Questions can be asked such as: 1) does library "one" cover the same diversity of chemical structures as library "two"; 2) if libraries "one" and "two" cover both different and identical aspects of diversity, how much overlap is there; 3) what about the possible overlap with libraries "three", "four", "five", etc.? To date, the prior art has been unable to answer these questions. In fact, assumptions have been made that as long as different chemistries were involved (ie., proteins, polysaccharides, small organic molecules), it was unlikely that the same diversity space was being sampled. However, such an assumption contradicts the well known reality that biological receptors can recognize molecular similarities arising from different structures. When screening for compounds possessing activity for undefined biological receptors, there is no way of telling a priori which chemistry or chemistries is most likely to produce molecules with activity for that receptor. Thus, screening with as many chemistries as possible is desired but is only really practical if redundant sampling of the same diversity space in each chemistry can be avoided. The prior art has not provided any guidance towards the resolution of these problems.

Brief Summary Of The Invention

In order to select a screening subset of a combinatorially accessible chemical universe which is representative of all the structural variation (diversity) to be found in the universe, it is necessary to have the means to describe and compare the molecular structural diversity in the universe. The first aspect of the present invention is the discovery of a generalized method of validating descriptors of molecular structural diversity. The method does not assume any prior knowledge of either the nature of the descriptor or of the biological system being studied and is generally applicable to all types of descriptors of molecular structure. This discovery enables several related advances to the art.

The second aspect of the invention is the discovery of a method of generating a validated three dimensional molecular structural descriptor using CoMFA fields. To generate these field descriptors required solving the alignment problem associated with these measurements. The alignment problem was solved using a topomeric procedure.

A third aspect of the invention is the discovery that validated molecular structural descriptors applicable to whole molecules can be used both to: 1) quantitatively define a meaningful end-point for selection in defining a single screening library (sampling procedure); and 2) merge libraries so as not to include molecules of the same or similar diversity. It is shown that a known metric (Tanimoto 2D fingerprint similarity) can be used in conjunction with the sampling procedure for this purpose.

A fourth aspect of the invention is the discovery of a method of using validated reactant and whole molecule molecular structural descriptors to rationally design a combinatorial screening library of optimal diversity. In particular, the shape sensitive topomeric CoMFA descriptor and the atom group Tanimoto 2D similarity descriptor may be used in the library design. As a benefit of designing a combinatorial screening library of optimal diversity based on validated molecular descriptors, many prior art problems associated with the synthesis, identification, and screening of mixtures of combinatorial molecules can be reduced or eliminated.

A fifth aspect of the invention is the use of validated molecular structural descriptors to guide the search for optimally active compounds after a lead compound has been identified by screening. In the case of a screening library designed for optimal diversity using validated descriptors, a great deal of the information necessary for lead optimization flows directly from the library design. In the case where a lead has been identified by screening a prior art library or through some other means, validated descriptors provide a method for identifying the molecular structural space nearest the lead which is most likely to contain compounds with the same or similar activity.

A sixth aspect of this invention is the discovery of a method for generating, using validated molecular descriptors, a virtual library of product molecules derivable from combinatorial reactions (or which may be represented by a combinatorial SLN [CSLN]) in which the characteristics of product molecules can be searched and compared without the actual construction of the product molecules. This virtual library allows the searching of billions of possible product molecules in reasonable amounts of time.

A seventh aspect of this invention is the discovery that, using validated molecular descriptors, the virtual library can be searched over billions of possible product molecules in ways to yield both optimally diverse screening libraries and to follow up on lead explosions. Using the virtual library, a much larger fraction of the chemically accessible universe can be searched for molecules of interest.

An eighth aspect of this invention is the discovery of a way to search, using validated molecular descriptors, the virtual library for possible molecules which have similar structures and/or activities to a query molecule which is not necessarily derived from a combinatorial synthesis. This discovery opens up a whole new method for seeking molecules with similar characteristics to a previously identified molecule.

It is an object of this invention to define a general process which may be used with

randomly selected literature data sets to validate molecular structural descriptors.

It is a further object of this invention to define a process to derive CoMFA steric fields (and, if desired, additional relevant fields) using topomeric alignment so that the resulting descriptor is valid.

It is a further object of this invention to teach that topomeric alignments may be used to describe molecular conformations.

It is a further object of this invention to define a general process for using a validated molecular descriptor to establish a meaningful end-point for the sampling of compounds thereby avoiding the oversampling of compounds representing the same molecular structural characteristics.

It is yet a further object of this invention to design an optimally diverse combinatorial screening library using multiple validated molecular structural descriptors.

It is a further object of this invention to use the topomeric CoMFA molecular structural descriptor as a reactant descriptor in the design of an optimally diverse combinatorial screening library.

It is a further object of this invention to use the Tanimoto 2D similarity molecular structural descriptor as a product descriptor in the design of an optimally diverse combinatorial screening library.

It is a further object of this invention to define a method for merging assemblies of molecules (libraries), both those designed by the methods of this invention and others not designed by the methods of this invention, in such a manner that molecules representing the same or similar diversity space are not likely to be included.

It is a further object of this invention to define methods for the use of validated molecular structural descriptors to guide the search for optimally active compounds after a lead compound has been identified by screening or some other method.

It is a further object of this invention to generate a virtual library, using validated molecular descriptors, of potential product molecules derivable from combinatorial reactions (or which may be represented by a combinatorial SLN [CSLN]) which can be searched for molecules having desired characteristics.

It is a further object of this invention to define methods for creating optimal diversity screening libraries as subsets of the virtual library.

It is still a further object of this invention to locate within the virtual library possible product molecules similar in structure and/or activity to lead compounds.

These and further objects of the invention will become apparent from the detailed description of the invention which follows.

Brief Description of Drawings

Figure 1 schematically shows the distribution of molecular structures around and about an island of biological activity in a hypothetical two dimensional metric space for a poorly designed prior art library and for an efficiently designed optimally diverse screening library.

Figure 2 shows a theoretical scatter plot (Patterson Plot) for a metric having the neighborhood property in which the X axis shows distances in some metric space calculated as the absolute value of the pairwise differences in some candidate molecular descriptor and the Y axis shows the absolute value of the pairwise differences in biological activity.

Figure 3 shows a Patterson plot for an illustrative data set.

Figure 4 shows a Patterson plot for the same data set as in Figure 3 but where the diversity descriptor values (X axis) associated with each molecule have been replaced by random numbers.

Figure 5 shows a Patterson plot for the same data set as in Figure 3 but where the diversity descriptor values (X axis) associated with each molecule have been replaced by a normalized force field strain energy/atom value.

Figure 6 shows three molecular structures numbered and marked in accordance with the topomeric alignment rule.

Figure 7 is a complete set of Patterson plots for the twenty data sets used for the validation studies of the topomeric CoMFA descriptor.

Figure 8 shows the two scatter plots displaying the relation between X^2 values and their corresponding density ratio values for the tested metrics over the twenty random data sets.

Figure 9 shows the graphs of the Tanimoto similarity measure vs. the pairwise frequency of active molecules for 18 groups examined from Index Chemicus.

Figure 10 shows a Patterson plot of the Cristalli data set using only those values which would have been used for a Tanimoto sigmoid plot of the same data set alongside a Patterson plot of the complete data set.

Figure 11 is a schematic of the combinatorial screening library design process.

Figure 12 shows a comparison of the volumes of space occupied by different molecules which are determined to be similar according to the Tanimoto 2D fingerprint descriptor but which are determined to be dissimilar according to the topomeric CoMFA field descriptor.

Figure 13 shows a plot of the Tanimoto 2D pairwise similarities for a typical combinatorial product universe.

Figure 14 shows the distribution of molecules resulting from a combinatorial screening library design plotted according to their Tanimoto 2D pairwise similarity after reactant filtering and after final product selection.

Figure 15 shows the distribution of molecules plotted according to their Tanimoto 2D pairwise similarity of three database libraries (Chapman & Hall) from the prior art.

Figure 16 shows a schematic representation of sets of possible reactants attached to a central core.

Figure 17 is a flowchart summarizing the overall process of virtual library construction.

Figures 18, 19, and 20 are a flowchart summarizing the overall process of applying the Tanimoto fingerprint metric for use in the virtual library.

Figures 21, 22, and 23 are a flowchart summarizing the overall process of using the Tanimoto fingerprint metric to search for molecules.

Figures 24, 25, and 26 are a flowchart summarizing the overall process of using both the topomeric CoMFA and Tanimoto metrics to search for molecules in the virtual library.

Figures 27, 28, 29, and 30 are a flowchart summarizing the overall process for topomeric searches of arbitrary query molecules.

Figure 31 shows the topomeric conformations of Tagamet and Zantac.

Disclosure Of Invention

1. Computational Chemistry Environment
2. Definitions
3. Validating Metrics
    - A. Theoretical Considerations - Neighborhood Property
    - B. Construction, Application, and Analysis Of Patterson Plots
4. Topomeric CoMFA Descriptor
    - A. Topomeric Alignment
        - i. General Topomeric Allignment
        - ii. Specialized Allignment for Chiral and Equivalent Atoms
    - B. Calculation Of CoMFA and Hydrogen Bonding Fields
    - C. Validation Of Topomeric CoMFA Descriptor
5. Tanimoto Fingerprint Descriptor

A. Neighborhood Property

B. Applicability Of Tanimoto To Different Biological Systems

C. Comparison of Sigmoid and Patterson Plots

6. Comparison of Tanimoto and Topomeric CoMFA Metrics

7. Additional Validation Results

8. Combinatorial Library Design Utilizing Validated Metrics

A. Removal Of Reactants For Non-Diversity Reasons

i. General Removal Criteria

ii. Biologically Based Criteria5

B. Removal of Non-Diverse Reactants

C. Identification (Building) Of Products

D. Removal Of Products For Non-Diversity Reasons

E. Removal of Non-Diverse Products

9. Lead Compound Optimization

A. Advantages Resulting From Product Filter

B. Advantages Resulting From Reactant Filter

C. Additional Optimization Methods Using Validated Metrics

10. Merging Libraries

11. Other Advantages of Optimally Diverse Libraries

12. Virtual Library Construction & Searching

A. Derivation of the Database (Virtual Library) of Compounds

B. Overview of Methodology

C. Overview of Virtual Library Construction

D. Virtual Library Construction

i. Representation of the Database of Compounds

ii. Application of A First Metric (Topomeric CoMFA)

iii. Application of A Second Metric (Tanimoto Fingerprint)

iv. Summary of Method & Scope of Chemistry

E. Searching the Virtual Library

i. Example Search Routine of Virtual Library - Tanimoto Similarity

ii. Design Screening Libraries (Subsets of the Virtual Library)

(a) Subset Screening Library Based On Topomeric Fields

and Tanimoto

(b) Subset Based on Tanimoto Similarity

(c) Subset Based on Topomeric Fields

(d) Subset Based on Combined Metric

iii. Designing Lead Optimizations

(a) Search Based on Tanimoto Similarity

(b) Searches Based on Topomer Similarity

(c) Topomeric (3D) Searching of Arbitrary Molecular Structures

(d) Topomeric (3D) Searching of Core Structures

1. Computational Chemistry Environment

Generally, all calculations and analyses to conduct combinatorial chemistry screening library design and follow up are implemented in a modern computational chemistry environment using software designed to handle molecular structures and associated properties and operations. For purposes of this Application, such an environment is specifically referenced. In particular, the computational environment and capabilities of the *SYBYL* and *UNITY* software programs developed and marketed by Tripos, Inc. (St. Louis, Missouri) are specifically utilized. Unless otherwise noted, all software references and commands in the following text are references to functionalities contained in the *SYBYL* and *UNITY* software programs. Where a required functionality is not available in *SYBYL* or *UNITY*, the software code to implement that functionality is provided in an Appendix to this Application. Software with similar functionalities to *SYBYL* and *UNITY* are available from other sources, both commercial and non-commercial, well known to those in the art. A general purpose programmable digital computer with ample amounts of memory and hard disk storage is required for the implementation of this invention. In performing the methods of this invention, representations of thousands of molecules and molecular structures as well as other data may need to be stored simultaneously in the random access memory of the computer or in rapidly available permanent storage. The inventors use a Silicon Graphics, Inc. *Challenge-M* computer having a single 150Mhz R4400 processor with 128 Mb memory and 4Gb hard disk storage space. As the size of the virtual library increases, a corresponding increase in hard disk storage and computational power is required. For these tasks, access to several gigabytes of storage and Silicon Graphics, Inc. processors in the R4400 to R10000 range are useful.

CLAIMS

What is claimed is:

1. A computer-based method for selecting, for all possible product molecules which could be created in a combinatorial synthesis from specified reactant molecules and common core molecule, a subset of product molecules, comprising the following steps:

a. Characterizing all the reactant molecules with a validated molecular structural descriptor appropriate to reactant molecules;

b. Hierarchically clustering the characterized reactant molecules until the intercluster distance corresponds to the neighborhood distance of the validated molecular structural descriptor or to a value close to the neighborhood distance which creates a logical clustering break;

c. Selecting a reactant molecule from each cluster;

d. Combinatorially assembling the selected reactant molecules and core molecule into products which would be created in the chemical synthesis;

e. Selecting a product molecule for inclusion in the subset;

f. Using a validated molecular structural descriptor appropriate to whole molecules, calculating the descriptor distance between all selected product molecules and all other product molecules;

g. Determining the shortest distance between each product molecule and all product molecules previously selected;

h. Selecting for inclusion in the subset the product molecule whose shortest descriptor distance from the previously selected molecules is the largest and is greater than the neighborhood distance of the descriptor;

i. Repeat steps f through h until the largest shortest difference between molecules is less than the neighborhood distance of the descriptor; and

j. Outputing a list of the selected product molecules and/or the reactant molecules from which the selected product molecules can be formed.

2. The method of claim 1 in which the validated molecular structural descriptor appropriate to reactant molecules is topomeric CoMFA fields.

3. The method of claim 2 in which topomeric hydrogen bond fields are used in conjunction with the topomeric CoMFA fields descriptor.

4. The method of claim 2 in which the validated molecular structural descriptor appropriate to whole molecules is the Tanimoto 2D coefficient.

5. The method of claim 4 in which before step a, reactant molecules with the following characteristics are removed from further use in the method:

a. toxic reactant molecules;

b. reactant molecules containing metals, improper forms of tautomers, and interfering chemical groups;

c. reactant molecules with too low a bioavailability;

d. reactant molecules not likely to cross membranes; and

e. reactant molecules containing biologically non-relevant groups.

6. The method of claim 5 in which before step e, product molecules with the following characteristics are removed from further use in the method:

a. product molecules having MW $\geq$ 750; and

b. product molecules not having a CLOGP between -2 and 7.5.

7. The method of claim 1 in which the validated molecular structural descriptor appropriate to whole molecules is the Tanimoto 2D coefficient.

8. The method of claim 7 in which before step a, reactant molecules with the following characteristics are removed from further use in the method:

a. toxic reactant molecules;

b. reactant molecules containing metals, improper forms of tautomers, and interfering chemical groups;

c. reactant molecules with too low a bioavailability;

d. reactant molecules not likely to cross membranes; and

e. reactant molecules containing biologically non-relevant groups.

9. The method of claim 8 in which before step e, product molecules with the following characteristics are removed from further use in the method:

a. product molecules having MW $\geq$ 750; and

b. product molecules not having a CLOGP between -2 and 7.5.

10. A computer-based method for selecting, for all possible product molecules which could be created in a combinatorial synthesis from specified reactant molecules, a subset of product molecules, comprising the following steps:

a. Characterizing all the reactant molecules with a validated molecular structural descriptor appropriate to reactant molecules;

b. Hierarchically clustering the characterized reactant molecules until the intercluster distance corresponds to the neighborhood distance of the validated molecular

structural descriptor or to a value close to the neighborhood distance which creates a logical clustering break;

c. Selecting a reactant molecule from each cluster;

d. Combinatorially assembling the selected reactant molecules and core molecule into products which would be created in the chemical synthesis;

e. Selecting a product molecule for inclusion in the subset;

f. Using a validated molecular structural descriptor appropriate to whole molecules, calculating the descriptor distance between all selected product molecules and all other product molecules;

g. Determining the shortest distance between each product molecule and all product molecules previously selected;

h. Selecting for inclusion in the subset the product molecule whose shortest descriptor distance from the previously selected molecules is the largest and is greater than the neighborhood distance of the descriptor;

i. Repeat steps f through h until the largest shortest difference between molecules is less than the neighborhood distance of the descriptor; and

j. Outputing a list of the selected product molecules and/or the reactant molecules from which the selected product molecules can be formed.

11. The method of claim 10 in which the validated molecular structural descriptor appropriate to reactant molecules is topomeric CoMFA fields.

12. The method of claim 11 in which topomeric hydrogen bond fields are used in conjunction with the topomeric CoMFA fields descriptor.

13. The method of claim 11 in which the validated molecular structural descriptor appropriate to whole molecules is the Tanimoto 2D coefficient.

14. The method of claim 13 in which before step a, reactant molecules with the following characteristics are removed from further use in the method:

a. toxic reactant molecules;

b. reactant molecules containing metals, improper forms of tautomers, and interfering chemical groups;

c. reactant molecules with too low a bioavailability;

d. reactant molecules not likely to cross membranes; and

e. reactant molecules containing biologically non-relevant groups.

15. The method of claim 14 in which before step e, product molecules with the following

characteristics are removed from further use in the method:

a. product molecules having MW $\geq$ 750; and

b. product molecules not having a CLOGP between -2 and 7.5.

16. The method of claim 10 in which the validated molecular structural descriptor appropriate to whole molecules is the Tanimoto 2D coefficient.

17. The method of claim 16 in which before step a, reactant molecules with the following characteristics are removed from further use in the method:

a. toxic reactant molecules;

b. reactant molecules containing metals, improper forms of tautomers, and interfering chemical groups;

c. reactant molecules with too low a bioavailability;

d. reactant molecules not likely to cross membranes; and

e. reactant molecules containing biologically non-relevant groups.

18. The method of claim 17 in which before step e, product molecules with the following characteristics are removed from further use in the method:

a. product molecules having MW $\geq$ 750; and

b. product molecules not having a CLOGP between -2 and 7.5.

19. A system for selecting, for all possible product molecules which can be created in a combinatorial synthesis from all specified reactant molecules and common core molecule, a subset of product molecules whose members collectively represent most of the molecular structural diversity in the possible combinatorially synthesized product molecules, comprising:

a. Means for characterizing all the reactant molecules with a validated molecular structural descriptor appropriate to reactant molecules;

b. Means for hierarchically clustering the characterized reactant molecules until the intercluster distance corresponds to the neighborhood distance of the validated molecular structural descriptor or to a value close to the neighborhood distance which creates a logical clustering break;

c. Means for selecting one reactant molecule from each cluster;

d. Means for combinatorially assembling the selected reactant molecules and core molecule into products which would be created in the chemical synthesis;

e. Means for selecting at least one product molecule for inclusion in the subset;

f. Means for using a validated molecular structural descriptor applicable to whole molecules for calculating the descriptor distance between all selected product

molecules and all other product molecules;

g. Means for determining the shortest distance between each product molecule and all product molecules previously selected;

h. Means for selecting for inclusion in the subset the product molecule whose shortest descriptor distance from the previously selected molecules is the largest and is greater than the neighborhood distance of the descriptor;

i. Means for invoking means f through h until the largest shortest difference between molecules is less than the neighborhood distance of the descriptor; and

j. Means for outputing a list of the selected product molecules and/or the reactant molecules from which the selected product molecules can be formed.

20. The system of claim 19 in which the reactant appropriate molecular structural descriptor is topomeric CoMFA fields.

21. The system of claim 20 in which topomeric hydrogen bond fields are used in conjunction with the topomeric CoMFA fields descriptor.

22. The system of claim 20 in which the whole molecule appropriate molecular structural descriptor is the Tanimoto 2D coefficient.

23. A system for selecting, for all possible product molecules which can be created in a combinatorial synthesis from all specified reactant molecules, a subset of product molecules whose members collectively represent most of the molecular structural diversity in the possible combinatorially synthesized product molecules, comprising:

a. Means for characterizing all the reactant molecules with a validated molecular structural descriptor appropriate to reactant molecules;

b. Means for hierarchically clustering the characterized reactant molecules until the intercluster distance corresponds to the neighborhood distance of the validated molecular structural descriptor or to a value close to the neighborhood distance which creates a logical clustering break;

c. Means for selecting one reactant molecule from each cluster;

d. Means for combinatorially assembling the selected reactant molecules into products which would be created in the chemical synthesis;

e. Means for selecting at least one product molecule for inclusion in the subset;

f. Means for using a validated molecular structural descriptor applicable to whole molecules for calculating the descriptor distance between all selected product molecules and all other product molecules;

g. Means for determining the shortest distance between each product molecule and all product molecules previously selected;

h. Means for selecting for inclusion in the subset the product molecule whose shortest descriptor distance from the previously selected molecules is the largest and is greater than the neighborhood distance of the descriptor;

i. Means for invoking means f through h until the largest shortest difference between molecules is less than the neighborhood distance of the descriptor; and

j. Means for outputing a list of the selected product molecules and/or the reactant molecules from which the selected product molecules can be formed.

24. The system of claim 23 in which the reactant appropriate molecular structural descriptor is topomeric CoMFA fields.

25. The system of claim 24 in which topomeric hydrogen bond fields are used in conjunction with the topomeric CoMFA fields descriptor.

26. The system of claim 24 in which the whole molecule appropriate molecular structural descriptor is the Tanimoto 2D coefficient.

27. A combinatorial screening library designed by a computer-based method, which selects the screening library molecules from those molecules which could be created in a combinatorial synthesis from specified reactant molecules and common core molecule, comprising the following steps:

a. Characterizing all the reactant molecules with a validated molecular structural descriptor appropriate to reactant molecules;

b. Hierarchically clustering the characterized reactant molecules until the intercluster distance corresponds to the neighborhood distance of the validated molecular structural descriptor or to a value close to the neighborhood distance which creates a logical clustering break;

c. Selecting a reactant molecule from each cluster;

d. Combinatorially assembling the selected reactant molecules and core molecule into products which would be created in the chemical synthesis;

e. Selecting a product molecule for inclusion in the subset;

f. Using a validated molecular structural descriptor appropriate to whole molecules, calculating the descriptor distance between all selected product molecules and all other product molecules;

g. Determining the shortest distance between each product molecule and all product molecules previously selected;

h. Selecting for inclusion in the subset the product molecule whose shortest descriptor distance from the previously selected molecules is the largest and is greater than the neighborhood distance of the descriptor;

i. Repeat steps f through h until the largest shortest difference between molecules is less than the neighborhood distance of the descriptor; and

j. Outputing a list of the selected product molecules and/or the reactant molecules from which the selected product molecules can be formed.

28. The method of claim 27 in which the validated molecular structural descriptor appropriate to reactant molecules is topomeric CoMFA fields.

29. The method of claim 28 in which topomeric hydrogen bond fields are used in conjunction with the topomeric CoMFA fields descriptor.

30. The method of claim 28 in which the validated molecular structural descriptor appropriate to whole molecules is the Tanimoto 2D coefficient.

31. A combinatorial screening library designed by a computer-based method, which selects the screening library molecules from those molecules which could be created in a combinatorial synthesis from specified reactant molecules, comprising the following steps:

a. Characterizing all the reactant molecules with a validated molecular structural descriptor appropriate to reactant molecules;

b. Hierarchically clustering the characterized reactant molecules until the intercluster distance corresponds to the neighborhood distance of the validated molecular structural descriptor or to a value close to the neighborhood distance which creates a logical clustering break;

c. Selecting a reactant molecule from each cluster;

d. Combinatorially assembling the selected reactant molecules and core molecule into products which would be created in the chemical synthesis;

e. Selecting a product molecule for inclusion in the subset;

f. Using a validated molecular structural descriptor appropriate to whole molecules, calculating the descriptor distance between all selected product molecules and all other product molecules;

g. Determining the shortest distance between each product molecule and all product molecules previously selected;

h. Selecting for inclusion in the subset the product molecule whose shortest descriptor distance from the previously selected molecules is the largest and is greater than the neighborhood distance of the descriptor;

i. Repeat steps f through h until the largest shortest difference between molecules is less than the neighborhood distance of the descriptor; and

j. Outputing a list of the selected product molecules and/or the reactant molecules from which the selected product molecules can be formed.

32. The method of claim 31 in which the validated molecular structural descriptor appropriate to reactant molecules is topomeric CoMFA fields.

33. The method of claim 32 in which topomeric hydrogen bond fields are used in conjunction with the topomeric CoMFA fields descriptor.

34. The method of claim 32 in which the validated molecular structural descriptor appropriate to whole molecules is the Tanimoto 2D coefficient.

35. A computer-based method for characterizing the relative validity or usefulness of molecular structural descriptors using multiple literature data sets containing a variety of chemical structures and associated activities comprising the following steps:

a. Applying the molecular structural descriptors to all compounds represented in each data set to derive descriptor values;

b. Constructing a Patterson plot for each molecular structural descriptor for each data set using the descriptor values for the compounds in each data set and their associated activities;

c. Determining the appropriate Patterson plot line and the corresponding density ratio for each molecular structural descriptor for each data set;

d. Determining the number of data sets for each molecular structural descriptor for which the Patterson plots have a density ratio greater than a predetermined cut-off value; and

e. Creating a ranking ratio for each molecular structural descriptor in which the numerator is the number determined in step d and the denominator is the number of data sets, said ranking ratio for each molecular structural descriptor being representative of the relative validity or usefulness of each molecular structural descriptor wherein higher values of the ranking ratio represent a higher degree of validity/usefulness.

36. The method of claim 35 in which in step d the predetermined cut-off is about 1.1.

37. A computer-based method of merging with a base assembly of molecules one or more additional assemblies of molecules, similar molecules in the assemblies having previously been identified and removed using a validated molecular structural descriptor, comprising the steps of:

a. Using a validated molecular structural descriptor which is appropriate to whole molecules, characterizing all the molecules in the base assembly of molecules and in the assembly of molecules to be merged;

b. Calculating the molecular structural distance between every molecule in the base assembly to every molecule in the assembly to be merged;

c. While there are still molecules in the assembly to be merged which have not been tested, selecting a molecule from the assembly to be merged;

d. Determining whether the molecular structural distance between the selected molecule and every molecule in the base assembly is within the neighborhood distance of the molecular structural descriptor;

e. Select for inclusion in the merged assemblies only those molecules identified in step d as having molecular structural distances greater than the neighborhood distance.

f. Repeat step c through step e until all molecules in the assembly to be merged have been tested; and

g. Repeat step a through step f for each additional assembly to be merged.

38. The method of claim 37 in which the molecular structural descriptor appropriate to whole molecules in the Tanimoto similarity coefficient.

39. A computer-based method of merging with a base assembly of molecules one or more additional assemblies of molecules, similar molecules in one or more of the assemblies having not previously been identified and removed using a validated molecular structural descriptor, comprising the steps of:

a. Selecting subsets of each assembly by:

(1) Selecting a molecule within each assembly;

(2) Using a validated molecular structural descriptor appropriate to whole molecules, calculating the descriptor distance between the selected molecule and all molecules within the assembly;

(3) Determining the shortest distance between the selected molecule and all

molecules previously selected for the subset;

(4) Selecting for inclusion in the subset the molecule whose shortest descriptor distance from the previously selected molecules is the largest and is greater than the neighborhood distance of the descriptor;

(5) Repeat steps (1) through (4) until the largest shortest difference between molecules is less than the neighborhood distance of the descriptor; and

(6) Repeat steps (1) through (5) for each assembly;

b. Using a validated molecular structural descriptor which is appropriate to whole molecules, characterizing all the molecules in the base assembly of molecules and in the assembly of molecules to be merged;

c. Calculating the molecular structural distance between every molecule in the base assembly to every molecule in the assembly to be merged;

d. While there are still molecules in the assembly to be merged which have not been tested, selecting a molecule from the assembly to be merged;

e. Determining whether the molecular structural distance between the selected molecule and every molecule in the base assembly is within the neighborhood distance of the molecular structural descriptor;

f. Select for inclusion in the merged assemblies only those molecules identified in step e as having molecular structural distances greater than the neighborhood distance.

g. Repeat step d through step f until all molecules in the assembly to be merged have been tested; and

h. Repeat step b through step g for each additional assembly to be merged.

40. The use of a subset of molecules, which could be made in a combinatorial synthesis of specified reactants and core, to specify the compounds to be synthesized and tested in biological screening assays, said subset being selected by the following computer-based method:

a. Characterizing all the reactant molecules with a validated molecular structural descriptor appropriate to reactant molecules;

b. Hierarchically clustering the characterized reactant molecules until the intercluster distance corresponds to the neighborhood distance of the validated molecular structural descriptor or to a value close to the neighborhood distance which creates a logical clustering break;

c. Selecting a reactant molecule from each cluster;

d. Combinatorially assembling the selected reactant molecules and core molecule into products which would be created in the chemical synthesis;

e. Selecting a product molecule for inclusion in the subset;

f. Using a validated molecular structural descriptor appropriate to whole molecules, calculating the descriptor distance between all selected product molecules and all other product molecules;

g. Determining the shortest distance between each product molecule and all product molecules previously selected;

h. Selecting for inclusion in the subset the product molecule whose shortest descriptor distance from the previously selected molecules is the largest and is greater than the neighborhood distance of the descriptor;

i. Repeat steps f through h until the largest shortest difference between molecules is less than the neighborhood distance of the descriptor; and

j. Outputing a list of the selected product molecules and/or the reactant molecules from which the selected product molecules can be formed.

41. The method of claim 40 in which the validated molecular structural descriptor appropriate to reactant molecules is topomeric CoMFA fields.

42. The method of claim 41 in which topomeric hydrogen bond fields are used in conjunction with the topomeric CoMFA fields descriptor.

43. The method of claim 41 in which the validated molecular structural descriptor appropriate to whole molecules is the Tanimoto 2D coefficient.

44. The molecules selected, from those which could be made in a combinatorial synthesis of specified reactants and core, by the following computer-based method:

a. Characterizing all the reactant molecules with a validated molecular structural descriptor appropriate to reactant molecules;

b. Hierarchically clustering the characterized reactant molecules until the intercluster distance corresponds to the neighborhood distance of the validated molecular structural descriptor or to a value close to the neighborhood distance which creates a logical clustering break;

c. Selecting a reactant molecule from each cluster;

d. Combinatorially assembling the selected reactant molecules and core molecule into products which would be created in the chemical synthesis;

e. Selecting a product molecule for inclusion in the subset;

f. Using a validated molecular structural descriptor appropriate to whole molecules, calculating the descriptor distance between all selected product molecules and all other product molecules;

g. Determining the shortest distance between each product molecule and all product molecules previously selected;

h. Selecting for inclusion in the subset the product molecule whose shortest descriptor distance from the previously selected molecules is the largest and is greater than the neighborhood distance of the descriptor;

i. Repeat steps f through h until the largest shortest difference between molecules is less than the neighborhood distance of the descriptor; and

j. Outputing a list of the selected product molecules and/or the reactant molecules from which the selected product molecules can be formed.

45. The method of claim 44 in which the validated molecular structural descriptor appropriate to reactant molecules is topomeric CoMFA fields.

46. The method of claim 45 in which topomeric hydrogen bond fields are used in conjunction with the topomeric CoMFA fields descriptor.

47. The method of claim 45 in which the validated molecular structural descriptor appropriate to whole molecules is the Tanimoto 2D coefficient.

48. A computer-based method of determining the neighborhood distance characteristic of a validated molecular structural descriptor using multiple literature data sets containing a variety of chemical structures and associated activities, comprising the following steps:

a. Applying the molecular structural descriptor to all compounds represented in each data set to derive descriptor values;

b. Constructing a Patterson plot for each molecular structural descriptor for each data set using the descriptor values for the compounds in each data set and their associated activities;

c. Determining the appropriate Patterson plot line for each data set;

d. Using for each data set a point on the Y axis of the corresponding Patterson plot the end point of an activity difference for which a neighborhood distance is desired, determining the X axis values of the molecular structural descriptor corresponding to the projection from the Patterson plot line of the end points of the activity difference;

e. Determining the average range of values for the neighborhood distance from the plots for each of the data sets.

49. A method of determining the molecules within any set which are most likely to have the same activity as a lead molecule previously identified in an assay comprising the following steps:

a. Characterizing the lead molecule and all other compounds to be examined using a validated molecular structural descriptor appropriate to whole molecules;

b. Determining the molecular structural descriptor distances between the lead molecule and all the other molecules; and

c. Identifying the molecules whose distances from the lead molecule fall within the neighborhood distance of the lead.

50. The method of claim 49 further comprising the additional steps of:

d. Determining the molecular structural descriptor distances between the set of molecules previously identified and all the other molecules excluding the lead and the sets;

e. Identifying the molecules whose distances from molecules in the previously selected set fall within the neighborhood distance; and

f. Repeating steps d through e as many times as desired.

51. A method of determining the useful boundaries of exploration within any set of molecular structures for molecules possessing the same activity as a lead molecule previously identified in an assay comprising the following steps:

a. Characterizing the lead molecule and all other compounds to be examined using a validated molecular structural descriptor appropriate to whole molecules;

b. Determining the molecular structural descriptor distances between the lead molecule and all the other molecules; and

c. Identifying the molecules whose distances from the lead molecule fall within the neighborhood distance of the lead;

d. Synthesizing and testing in an assay the molecules identified in step c and if no activity is detected, stop.

e. If activity is detected, calculating molecular structural descriptor distances, from each molecule identified in the previous step as showing activity, to all other compounds (excluding the lead compound and each previously identified active compound);

f. Identifying all molecules within the neighborhood diameter of the previously identified active molecules;

g. Synthesizing and testing in an assay the molecules identified in the previous step, and

if no activity is detected, stop; and

h. Repeating steps e through g until no further compounds show activity in the assay.

52. A computer-based method of characterizing the three dimensional structure of reactants, which can assume many conformations, comprising the steps of:

a. Topomerically aligning the reactants; and

b. Determining the CoMFA steric fields for each topomerically aligned reactant.

53. The method of claim 52 further comprising the addition of topomeric hydrogen bonding fields to the CoMFA steric fields.

54. A computer-based method of applying a molecular structural descriptor to a set of reactants comprising the following steps:

a. Topomerically aligning the reactants;

b. Determining the CoMFA steric fields for each topomerically aligned reactant; and

c. Calculating the field differences between all pairs of reactants.

55. The method of claim 54 further comprising after step b the additional step of adding topomeric hydrogen bonding fields to the CoMFA fields.

56. The method of claim 54 further comprising after step c the additional step of hierarchically clustering the reactants until the intercluster distance is about 80 - 100 CoMFA field units.

57. In a digital computer in which representations of specified reactant molecules and a core molecule have been stored, a computer-based method for selecting, for all possible product molecules which could be created in a combinatorial synthesis from the reactant molecules and common core molecule, a subset of product molecules, comprising the following steps:

a. Characterizing all the reactant molecules with a validated molecular structural descriptor appropriate to reactant molecules;

b. Hierarchically clustering the characterized reactant molecules until the intercluster distance corresponds to the neighborhood distance of the validated molecular structural descriptor or to a value close to the neighborhood distance which creates a logical clustering break;

c. Selecting a reactant molecule from each cluster;

d. Combinatorially assembling the selected reactant molecules and core molecule into products which would be created in the chemical synthesis;

e. Selecting a product molecule for inclusion in the subset;

f. Using a validated molecular structural descriptor appropriate to whole molecules, calculating the descriptor distance between all selected product molecules and all other product molecules;

g. Determining the shortest distance between each product molecule and all product molecules previously selected;

h. Selecting for inclusion in the subset the product molecule whose shortest descriptor distance from the previously selected molecules is the largest and is greater than the neighborhood distance of the descriptor;

i. Repeat steps f through h until the largest shortest difference between molecules is less than the neighborhood distance of the descriptor; and

j. Outputing a list of the selected product molecules and/or the reactant molecules from which the selected product molecules can be formed.

58. The method of claim 57 in which the validated molecular structural descriptor appropriate to reactant molecules is topomeric CoMFA fields.

59. The method of claim 58 in which topomeric hydrogen bond fields are used in conjunction with the topomeric CoMFA fields descriptor.

60. The method of claim 57 in which the validated molecular structural descriptor appropriate to whole molecules is the Tanimoto 2D coefficient.

61. A computer-based method for generating a virtual library of possible combinatorially derived product molecules which can be searched for product molecules having desired properties without the necessity of generating the product structures during the search, comprising the following steps:

a. Creating one or more files identifying one or more combinatorial reactions for one or more core structures;

b. Creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

c. Associating with each structural variation, data, characterizing each structural variation including:

(1) Characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(2) Characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations.

62. A virtual library of possible combinatorially derived product molecules which can be searched for product molecules having desired properties without the necessity of generating the product structures during the search, generated by the following process:

a. Creating one or more files identifying one or more combinatorial reactions for one or more core structures;

b. Creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

c. Associating with each structural variation, data, characterizing each structural variation including:

(1) Characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(2) Characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations.

63. The method of claim 61 further comprising a computer-based method for selecting from the virtual library, for all possible product molecules which could be created by all combinatorial arrangements of specified structural variations and a common core molecule, a subset of product molecules, comprising the following additional steps:

b. identifying all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

c. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

d. using a validated molecular descriptor appropriate to whole molecules with which the Virtual Library was generated, removing from the set of all remaining molecules those molecules falling within a chosen neighborhood distance of the selected molecule;

e. using a validated molecular descriptor appropriate to the structural variations with which the Virtual Library was generated, removing from the set of all remaining product molecules those molecules formed from structural variations falling within a chosen neighborhood distance of the structural variations of the selected molecule;

f. selecting from the set of all product molecules remaining after step e a product molecule for inclusion in the subset;

g. repeating steps d through f until no additional product molecules remain to be selected in step f; and

h. Outputting a list of the selected subset and/or the structural variations from which the subset can be formed.

64. The method of claim 61 further comprising a computer-based method for selecting from the virtual library, for all possible product molecules which could be created by all combinatorial arrangements of specified structural variations and core molecules, a subset of product molecules, comprising the following additional steps:

b. selecting from all possible cores a core upon which to base the subset;

c. using a validated molecular descriptor appropriate to cores, selecting from the set of all possible cores those core molecules falling within the neighborhood distance of the selected core molecule;

d. identifying all possible combinatorial product molecules which could result from the specified structural variations and selected core molecules;

e. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

f. using a validated molecular descriptor appropriate to whole molecules with which the Virtual Library was generated, removing from the set of all remaining molecules those molecules falling within a chosen neighborhood distance of the selected molecule;

g. using a validated molecular descriptor appropriate to the structural variations with which the Virtual Library was generated, removing from the set of all remaining product molecules those molecules formed from structural variations falling within a chosen neighborhood distance of the structural variations of the selected molecule;

h. selecting from the set of all product molecules remaining after step g a product molecule for inclusion in the subset;

i. repeating steps f through h until no additional product molecules remain to be selected in step h; and

j. Outputting a list of the selected subset and/or the structural variations and cores from which the subset can be formed.

65. The method of claim 61 further comprising a computer-based method for selecting from the virtual library, for all possible product molecules which could be created by all combinatorial arrangements of specified structural variations and a common core molecule, a subset of product molecules, comprising the following additional steps:

b. identifying all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

c. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

d. using a validated molecular descriptor appropriate to whole molecules with which the Virtual Library was generated, removing from the set of all remaining molecules those molecules falling within the neighborhood distance of the selected molecule;

e. selecting from the set of all product molecules remaining after step **d** a product molecule for inclusion in the subset;

f. repeating steps **d** through **e** until no additional product molecules remain to be selected in step **f**; and

g. Ouputting a list of the selected subset and/or the structural variations from which the subset can be formed.

66. The method of claim 61 further comprising a computer-based method for selecting from the virtual library, for all possible product molecules which could be created by all combinatorial arrangements of specified structural variations and a common core molecule, a subset of product molecules, comprising the following additional steps:

b. identifying all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

c. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

d. using a validated molecular descriptor appropriate to the structural variations with which the Virtual Library was generated, removing from the set of all remaining product molecules those molecules formed from structural variations falling within a chosen neighborhood distance of the structural variations of the selected molecule;

e. selecting from the set of all product molecules remaining after step **d** a product molecule for inclusion in the subset;

f. repeating steps d through e until no additional product molecules remain to be selected in step e; and

g. Ouputting a list of the selected subset and/or the structural variations from which the subset can be formed.

67. A screening library designed by a computer-based method which selects the screening library molecules from those molecules which could be created by all combinatorial arrangements of specified structural variations and a common core molecule comprising the following steps:

a. generating a virtual library by:

(1). creating one or more files identifying one or more combinatorial reactions for one or more core structures;

(2). creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

(3). associating with each structural variation, data, characterizing each structural variation including:

(a). characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(b). characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations;

b. identifying in the virtual library all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

c. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

d. using a validated molecular descriptor appropriate to whole molecules with which the Virtual Library was generated, removing from the set of all remaining molecules those molecules falling within a chosen neighborhood distance of the selected molecule;

e. using a validated molecular descriptor appropriate to the structural variations with which the Virtual Library was generated, removing from the set of all remaining product

molecules those molecules formed from structural variations falling within a chosen neighborhood distance of the structural variations of the selected molecule;

f. selecting from the set of all product molecules remaining after step e a product molecule for inclusion in the subset;

g. repeating steps d through f until no additional product molecules remain to be selected in step f; and

h. Outputting a list of the selected subset and/or the structural variations from which the subset can be formed.

68. A screening library designed by a computer-based method which selects the screening library molecules from those molecules which could be created by all combinatorial arrangements of specified structural variations and core molecules comprising the following steps:

a. generating a virtual library by:

(1). creating one or more files identifying one or more combinatorial reactions for one or more core structures;

(2). creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

(3). associating with each structural variation, data, characterizing each structural variation including:

(a). characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(b). characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations;

b. selecting from all possible cores a core upon which to base the subset;

c. using a validated molecular descriptor appropriate to cores, selecting from the set of all possible cores those core molecules falling within the neighborhood distance of the selected core molecule;

d. identifying all possible combinatorial product molecules which could result from the

specified reactants and selected core molecules;

e. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

f. using a validated molecular descriptor appropriate to whole molecules with which the Virtual Library was generated, removing from the set of all remaining molecules those molecules falling within a chosen neighborhood distance of the selected molecule;

g. using a validated molecular descriptor appropriate to the structural variations with which the Virtual Library was generated, removing from the set of all remaining product molecules those molecules formed from structural variations falling within a chosen neighborhood distance of the structural variations of the selected molecule;

h. selecting from the set of all product molecules remaining after step g a product molecule for inclusion in the subset;

i. repeating steps f through h until no additional product molecules remain to be selected in step h; and

j. Outputting a list of the selected subset and/or the structural variations and cores from which the subset can be formed.

69. The use of a subset of molecules, which could be made in a combinatorial synthesis of specified reactants and common core, to specify the compounds to be synthesized and tested in appropriate assays, said subset being selected by the following computer-based method:

a. generating a virtual library by:

(1). creating one or more files identifying one or more combinatorial reactions for one or more core structures;

(2). creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

(3). associating with each structural variation, data, characterizing each structural variation including:

(a). characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(b). characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed

combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations;

b. identifying in the virtual library all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

c. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

d. using a validated molecular descriptor appropriate to whole molecules with which the Virtual Library was generated, removing from the set of all remaining molecules those molecules falling within a chosen neighborhood distance of the selected molecule;

e. using a validated molecular descriptor appropriate to the structural variations with which the Virtual Library was generated, removing from the set of all remaining product molecules those molecules formed from structural variations falling within a chosen neighborhood distance of the structural variations of the selected molecule;

f. selecting from the set of all product molecules remaining after step e a product molecule for inclusion in the subset;

g. repeating steps d through f until no additional product molecules remain to be selected in step f; and

h. Outputting a list of the selected subset and/or the reactants from which the subset can be formed.

70. The molecules selected, from those which could be made in a combinatorial synthesis of specified reactants and common core, by the following computer-based method:

a. generating a virtual library by:

(1). creating one or more files identifying one or more combinatorial reactions for one or more core structures;

(2). creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

(3). associating with each structural variation, data, characterizing each structural variation including:

(a). characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(b). characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations;

b. identifying in the virtual library all possible combinatorial product molecules which could result from the specified reactants and core molecule;

c. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

d. using a validated molecular descriptor appropriate to whole molecules with which the Virtual Library was generated, removing from the set of all remaining molecules those molecules falling within a chosen neighborhood distance of the selected molecule;

e. using a validated molecular descriptor appropriate to the structural variations with which the Virtual Library was generated, removing from the set of all remaining product molecules those molecules formed from structural variations falling within a chosen neighborhood distance of the structural variations of the selected molecule;

f. selecting from the set of all product molecules remaining after step e a product molecule for inclusion in the subset;

g. repeating steps d through f until no additional product molecules remain to be selected in step f; and

h. Outputting a list of the selected subset and/or the reactants from which the subset can be formed.

71. The molecules selected, from those which could be made in a combinatorial synthesis of specified reactants and cores, by the following computer-based method:

a. generating a virtual library by:

(1). creating one or more files identifying one or more combinatorial reactions for one or more core structures;

(2). creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

(3). associating with each structural variation, data, characterizing each structural variation including:

(a). characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed

combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(b). characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations;

b. selecting from all possible cores a core upon which to base the subset;

c. using a validated molecular descriptor appropriate to cores, selecting from the set of all possible cores those core molecules falling within the neighborhood distance of the selected core molecule;

d. identifying all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

e. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

f. using a validated molecular descriptor appropriate to whole molecules with which the Virtual Library was generated, removing from the set of all remaining molecules those molecules falling within a chosen neighborhood distance of the selected molecule;

g. using a validated molecular descriptor appropriate to the structural variations with which the Virtual Library was generated, removing from the set of all remaining product molecules those molecules formed from structural variations falling within a chosen neighborhood distance of the structural variations of the selected molecule;

h. selecting from the set of all product molecules remaining after step g a product molecule for inclusion in the subset;

i. repeating steps f through h until no additional product molecules remain to be selected in step h; and

j. Outputting a list of the selected subset and/or the reactants from which the subset can be formed.

72. The method of claim 1 further comprising a computer-based method for selecting from the virtual library, for all possible product molecules which could be created by all combinatorial arrangements of specified structural variations and a common core molecule, a subset of product molecules, comprising the following additional steps:

b. identifying all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

c. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

d. using a combination validated molecular descriptor characterizing both whole molecule and structural variation features with which the Virtual Library was generated, removing from the set of all remaining molecules those molecules falling within a chosen neighborhood distance of the selected molecule;

e. selecting from the set of all product molecules remaining after step d a product molecule for inclusion in the subset;

f. repeating steps d through e until no additional product molecules remain to be selected in step e; and

h. Outputting a list of the selected subset and/or the structural variations from which the subset can be formed.

73. The method of claim 61 further comprising a computer-based method for selecting from the virtual library, for all possible product molecules which could be created by all combinatorial arrangements of specified structural variations and core molecules, a subset of product molecules, comprising the following additional steps:

b. selecting from all possible cores a core upon which to base the subset;

c. using a validated molecular descriptor appropriate to cores, selecting from the set of all possible cores those core molecules falling within the neighborhood distance of the selected core molecule;

d. identifying all possible combinatorial product molecules which could result from the specified structural variations and selected core molecules;

e. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

f. using a combination validated molecular descriptor characterizing both whole molecule and structural variation features with which the Virtual Library was generated, removing from the set of all remaining molecules those molecules falling within a chosen neighborhood distance of the selected molecule;

g. selecting from the set of all product molecules remaining after step e a product molecule for inclusion in the subset;

f. repeating steps e through g until no additional product molecules remain to be selected in step g; and

h. Outputting a list of the selected subset and/or the structural variations and cores from

which the subset can be formed.

74. The molecules selected, from those which could be made in a combinatorial synthesis of specified reactants and common core, by the following computer-based method:

a. generating a virtual library by:

(1). creating one or more files identifying one or more combinatorial reactions for one or more core structures;

(2). creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

(3). associating with each structural variation, data, characterizing each structural variation including:

(a). characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(b). characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations;

b. identifying in the virtual library all possible combinatorial product molecules which could result from the specified reactants and core molecule;

c. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

d. using a combination validated molecular descriptor characterizing both whole molecule and structural variation features with which the Virtual Library was generated, removing from the set of all remaining molecules those molecules falling within a chosen neighborhood distance of the selected molecule;

e. selecting from the set of all product molecules remaining after step **d** a product molecule for inclusion in the subset;

f. repeating steps **d** through **e** until no additional product molecules remain to be selected in step e; and

h. Outputting a list of the selected subset and/or the reactants from which the subset can be formed.

75. The molecules selected, from those which could be made in a combinatorial synthesis of specified reactants and cores, by the following computer-based method:

a. generating a virtual library by:

(1). creating one or more files identifying one or more combinatorial reactions for one or more core structures;

(2). creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

(3). associating with each structural variation, data, characterizing each structural variation including:

(a). characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(b). characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations;

b. selecting from all possible cores a core upon which to base the subset;

c. using a validated molecular descriptor appropriate to cores, selecting from the set of all possible cores those core molecules falling within the neighborhood distance of the selected core molecule;

d. identifying all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

e. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

f. using a combination validated molecular descriptor characterizing both whole molecule and structural variation features with which the Virtual Library was generated, removing from the set of all remaining molecules those molecules falling within a chosen neighborhood distance of the selected molecule;

g. selecting from the set of all product molecules remaining after step f a product molecule for inclusion in the subset;

f. repeating steps f through g until no additional product molecules remain to be selected

in step g; and

h. Outputting a list of the selected subset and/or the reactants and cores from which the subset can be formed.

76. The method of claim 61 further comprising a method of determining within the virtual library, the molecules which could be created by all combinatorial arrangements of specified structural variations and a common core molecule, which are most likely to have the same type of activity as a molecule of interest comprising the following steps:

a. identifying in the virtual library all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

b. characterizing the molecule of interest with a validated molecular structural descriptor appropriate to whole molecules with which the virtual library was generated;

d. using the same validated molecular descriptor appropriate to whole molecules, selecting the set of all possible molecules whose descriptor values fall within a chosen neighborhood distance of the selected molecule; and

g. Ouputting a list of the selected subset and/or the structural variations from which the subset can be formed.

77. The method of claim 61 further comprising a method of determining within the virtual library, the molecules which could be created by all combinatorial arrangements of specified structural variations and a common core molecule, which are most likely to have the same type of activity as a molecule of interest comprising the following steps:

a. identifying in the virtual library all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

b. characterizing the molecule of interest with a validated molecular structural descriptor appropriate to structural variations with which the virtual library was generated;

d. using the same validated molecular descriptor appropriate to structural variations, selecting the set of all possible molecules whose descriptor values fall within a chosen neighborhood distance of the selected molecule; and

g. Ouputting a list of the selected subset and/or the structural variations from which the subset can be formed.

78. The method of claim 61 further comprising a method of determining within the virtual library, the molecules which could be created by all combinatorial arrangements of specified structural variations and a common core molecule, which are most likely to have the same type of activity as a molecule of interest comprising the following steps:

a. identifying in the virtual library all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

b. characterizing the molecule of interest with both a validated molecular structural descriptor appropriate to structural variations with which the virtual library was generated and with a validated molecular structural descriptor appropriate to structural variations with which the virtual library was generated;

d. using the same validated molecular descriptor appropriate to whole molecules, selecting the set of all possible molecules whose descriptor values fall within a chosen neighborhood distance of the selected molecule, and using the same validated molecular descriptor appropriate to structural variations, selecting the set of all possible molecules whose descriptor values fall within a chosen neighborhood distance of the selected molecule; and

e. Ouputting a list of the selected subset and/or the structural variations from which the subset can be formed.

79. The method of claim 61 further comprising a method of determining within the virtual library, the molecules which could be created by all combinatorial arrangements of specified structural variations and a common core molecule, which are most likely to have the same type of activity as a molecule of interest comprising the following steps:

a. identifying in the virtual library all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

b. characterizing the molecule of interest with a combination validated molecular descriptor, characterizing both whole molecule and structural variation features, with which the Virtual Library was generated;

d. using the same validated molecular descriptor, selecting the set of all possible molecules whose descriptor values fall within a chosen neighborhood distance of the selected molecule; and

g. Ouputting a list of the selected subset and/or the structural variations from which the subset can be formed.

80. The molecules, which are most likely to have the same type of activity as a molecule of interest, selected, from those which could be made in a combinatorial synthesis from specified reactants and a common core molecule, by the following computer-based method:

a. generating a virtual library by:

(1). creating one or more files identifying one or more combinatorial reactions for one

or more core structures;

(2). creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

(3). associating with each structural variation, data, characterizing each structural variation including:

(a). characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(b). characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations;

b. identifying in the virtual library all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

c. characterizing the molecule of interest with both a validated molecular structural descriptor appropriate to structural variations with which the virtual library was generated and with a validated molecular structural descriptor appropriate to structural variations with which the virtual library was generated;

d. using the same validated molecular descriptor appropriate to whole molecules, selecting the set of all possible molecules whose descriptor values fall within a chosen neighborhood distance of the selected molecule, and using the same validated molecular descriptor appropriate to structural variations, selecting the set of all possible molecules whose descriptor values fall within a chosen neighborhood distance of the selected molecule; and

e. Ouputting a list of the selected subset and/or the reactants from which the subset can be formed.

81. The molecules, which are most likely to have the same type of activity as a molecule of interest, selected, from those which could be made in a combinatorial synthesis from specified reactants and a common core molecule, by the following computer-based method:

a. generating a virtual library by:

(1). creating one or more files identifying one or more combinatorial reactions for one

or more core structures;

(2). creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

(3). associating with each structural variation, data, characterizing each structural variation including:

(a). characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(b). characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations;

b. identifying in the virtual library all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

c. characterizing the molecule of interest with a combination validated molecular descriptor, characterizing both whole molecule and structural variation features, with which the Virtual Library was generated;

d. using the same validated molecular descriptor, selecting the set of all possible molecules whose descriptor values fall within a chosen neighborhood distance of the selected molecule; and

e. Ouputting a list of the selected subset and/or the reactant from which the subset of molecules can be formed.

82. The use of a subset of molecules, which are most likely to have the same type of activity as a molecule of interest and selected from those which could be made in a combinatorial synthesis from specified reactants and a common core molecule, to specify the compounds to be synthesized and tested in appropriate assays, said subset being selected by the following computer-based method:

a. generating a virtual library by:

(1). creating one or more files identifying one or more combinatorial reactions for one or more core structures;

(2). creating separate structural variation files (associated with the reaction identifying

files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

(3). associating with each structural variation, data, characterizing each structural variation including:

(a). characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(b). characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations;

b. identifying in the virtual library all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

c. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

d. characterizing the molecule of interest with both a validated molecular structural descriptor appropriate to whole molecules with which the virtual library was generated and with a validated molecular structural descriptor appropriate to structural variations with which the virtual library was generated;

e. using the same validated molecular descriptor appropriate to whole molecules, selecting the set of all possible molecules whose descriptor values fall within a chosen neighborhood distance of the selected molecule, and using the same validated molecular descriptor appropriate to structural variations, selecting the set of all possible molecules whose descriptor values fall within a chosen neighborhood distance of the selected molecule; and

f. Ouputting a list of the selected subset and/or the reactants from which the subset can be formed.

83. The use of a subset of molecules, which are most likely to have the same type of activity as a molecule of interest and selected from those which could be made in a combinatorial synthesis from specified reactants and a common core molecule, to specify the compounds to be synthesized and tested in appropriate assays, said subset being selected by the following computer-based method:

a. generating a virtual library by:

(1). creating one or more files identifying one or more combinatorial reactions for one or more core structures;

(2). creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

(3). associating with each structural variation, data, characterizing each structural variation including:

(a). characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(b). characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations;

b. identifying in the virtual library all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

c. selecting from all possible combinatorial product molecules a product molecule for inclusion in the subset;

d. characterizing the molecule of interest with a combination validated molecular descriptor, characterizing both whole molecule and structural variation features, with which the Virtual Library was generated;

e. using the same validated molecular descriptor, selecting the set of all possible molecules whose descriptor values fall within a chosen neighborhood distance of the selected molecule; and

f. Ouputting a list of the selected subset and/or the reactant from which the subset of molecules can be formed.

84. The method of claim 61 further comprising a method of determining within the virtual library, the molecules which could be created by all combinatorial arrangements of specified structural variations and core molecules, which are most likely to have the same type of activity as a molecule of interest, comprising the following steps:

a. selecting from all possible cores a core upon which to base the subset;

b. using a validated molecular descriptor appropriate to cores, selecting from the set of all possible cores those core molecules falling within the neighborhood distance of the selected core molecule;

c. identifying all possible combinatorial product molecules which could result from the specified reactants and selected core molecules;

d. selecting and characterizing the molecule of interest with a validated molecular structural descriptor appropriate to whole molecules with which the virtual library was generated;

e. using the same validated molecular descriptor appropriate to whole molecules, selecting the set of all possible molecules whose descriptor values fall within a chosen neighborhood distance of the selected molecule; and

f. Ouputting a list of the selected subset and/or the structural variations from which the subset can be formed.

85. The method of claim 61 further comprising a method of determining within the virtual library, the molecules which could be created by all combinatorial arrangements of structural variations and core molecules, which are most likely to have the same type of activity as a molecule of interest, which is not known to be derived from a combinatorial reaction, comprising the following steps:

a. fragmenting the molecule of interest as described in a fragmentation table;

b. selecting a fragmentation pattern;

c. aligning the fragments according to topomeric alignment rules;

d. generating CoMFA fields for each aligned fragment;

e. identifying which reaction types within the virtual library correspond to the reaction type resulting from the fragmentation;

f. identifying whether the fragmentation pattern generated a core, and, if so, implementing the following steps:

(1) characterizing the core with CoMFA fields; and

(2) identifying, by comparing the field values, whether the core resembles any cores used in the creation of the virtual library;

g. selecting structural variations which were used in generating the virtual library with cores which matched the core resulting from the fragmentation;

h. comparing the CoMFA fields of the topomerically aligned fragments with the fields of the identified structural variations by taking the root sum of squares field differences;

i. selecting those structural variations for which the root sum of squares field difference

falls within a chosen neighborhood value;

j. ouputting a list of the selected subset and/or the structural variations from which the subset can be forme;

k. repeating steps b through j for all possible fragments.

86. The molecules, which are most likely to have the same type of activity as a molecule of interest which is not known to be derived from a combinatorial reaction, selected from those product molecules which could be created by all combinatorial arrangements of structural variations and core molecules, by the following computer-based method:

a. generating a virtual library by:

(1). creating one or more files identifying one or more combinatorial reactions for one or more core structures;

(2). creating separate structural variation files (associated with the reaction identifying files) in which are listed together the structural variations representative of those reactants which will react at each variation site of each combinatorial reaction;

(3). associating with each structural variation, data, characterizing each structural variation including:

(a). characterization data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has not been derived from the application of validated molecular structural descriptors; and

(b). characterizing data, taking into account when necessary the structures of the cores with which the structural variations would be combined in the listed combinatorial syntheses, which has been derived from applying validated molecular structural descriptors to the structural variations;

b. fragmenting the molecule of interest as described in a fragmentation table;

c. selecting a fragmentation pattern;

d. aligning the fragments according to topomeric alignment rules;

e. generating CoMFA fields for each aligned fragment;

f. identifying which reaction types within the virtual library correspond to the reaction type resulting from the fragmentation;

g. identifying whether the fragmentation pattern generated a core, and, if so, implementing the following steps:

(1) characterizing the core with CoMFA fields; and

(2) identifying, by comparing the field values, whether the core resembles any cores used in the creation of the virtual library;

h. selecting structural variations which were used in generating the virtual library with cores which matched the core resulting from the fragmentation;

i. comparing the CoMFA fields of the topomerically aligned fragments with the fields of the identified structural variations by taking the root sum of squares field differences;

j. selecting those structural variations for which the root sum of squares field difference falls within a chosen neighborhood value;

k. ouputting a list of the selected subset and/or the structural variations from which the subset can be forme;

l. repeating steps c through k for all possible fragments.

87. The method of claims 63 or 65 or 69 or 71 or 72 or 73 or 74 or 75 or 80 or 86 or 88 in which the following additional step is performed immediately after the step of using a validated molecular descriptor appropriate to whole molecules:

t. repeating the previous step for another validated molecular descriptor appropriate to whole molecules with which the Virtual Library was generated until no additional whole molecule descriptor remains to be used.

88. The method of claims 63 or 65 or 70 or 71 or 72 or 73 or 74 or 75 or 81 or 86 in which the following additional step is performed immediately after the step of using a validated molecular descriptor appropriate to structural variations:

u. repeating the previous step for another validated molecular descriptor appropriate to structural variations with which the Virtual Library was generated until no additional structural variation descriptor remains to be used.

89. The method of claim 63 in which the additional step t is performed immediately after the step of using a validated molecular descriptor appropriate to whole molecules and further in which step u is performed immediately after the step of using a validated molecular descriptor appropriate to structural variations:

t. repeating the previous step for another validated molecular descriptor appropriate to whole molecules with which the Virtual Library was generated until no additional whole molecule descriptor remains to be used; and

u. repeating the previous step for another validated molecular descriptor appropriate to structural variations with which the Virtual Library was generated until no additional structural variation descriptor remains to be used.

90. The method of claims 61 or 63 or 65 or 70 or 71 or 72 or 73 or 74 or 86 in which the validated molecular structural descriptor appropriate to structural variations is topomeric CoMFA fields.

91. The method of claim 61 or 63 or 65 or 70 or 71 or 72 or 73 or 74 or 86 in which topomeric hydrogen bond fields are used in conjunction with the topomeric CoMFA fields descriptor.

92. The method of claims 63 or 65 or 69 or 71 or 72 or 73 or 74 or 75 or 80 or 86 or 88 in which the validated molecular structural descriptor appropriate to whole molecules is the Tanimoto 2D coefficient.

93. The method of claim 63 in which after step g product molecules with the following characteristics are removed from further use in the method:

a. toxic reactant molecules;

b. reactant molecules containing metals, improper forms of tautomers, and interfering chemical groups;

c. reactant molecules with too low a bioavailability;

d. reactant molecules not likely to cross membranes; and

e. reactant molecules containing biologically non-relevant groups.

94. The method of claim 63 in which after step g product molecules with the following characteristics are removed from further use in the method:

a. product molecules having MW $\geq$ 750; and

b. product molecules not having a CLOGP between -2 and 7.5.

95. The methods of selecting screening libraries as disclosed in this invention.

96. The systems for selecting screening libraries as disclosed in this invention.

97. The screening libraries selected by the methods or systems disclosed in this invention.

98. The metric validation method as disclosed in this invention.

99. The method of merging libraries as disclosed in this invention.

100. The method of lead explosion as disclosed in this invention.

101. The methods of molecular alignment as disclosed in this invention.

102. The new molecular structural descriptors as disclosed in this invention.

103. The methods of generating a virtual library as disclosed in this invention.

104. The methods of searching a virtual library as disclosed in this invention.

105. The virtual library as disclosed in this invention.

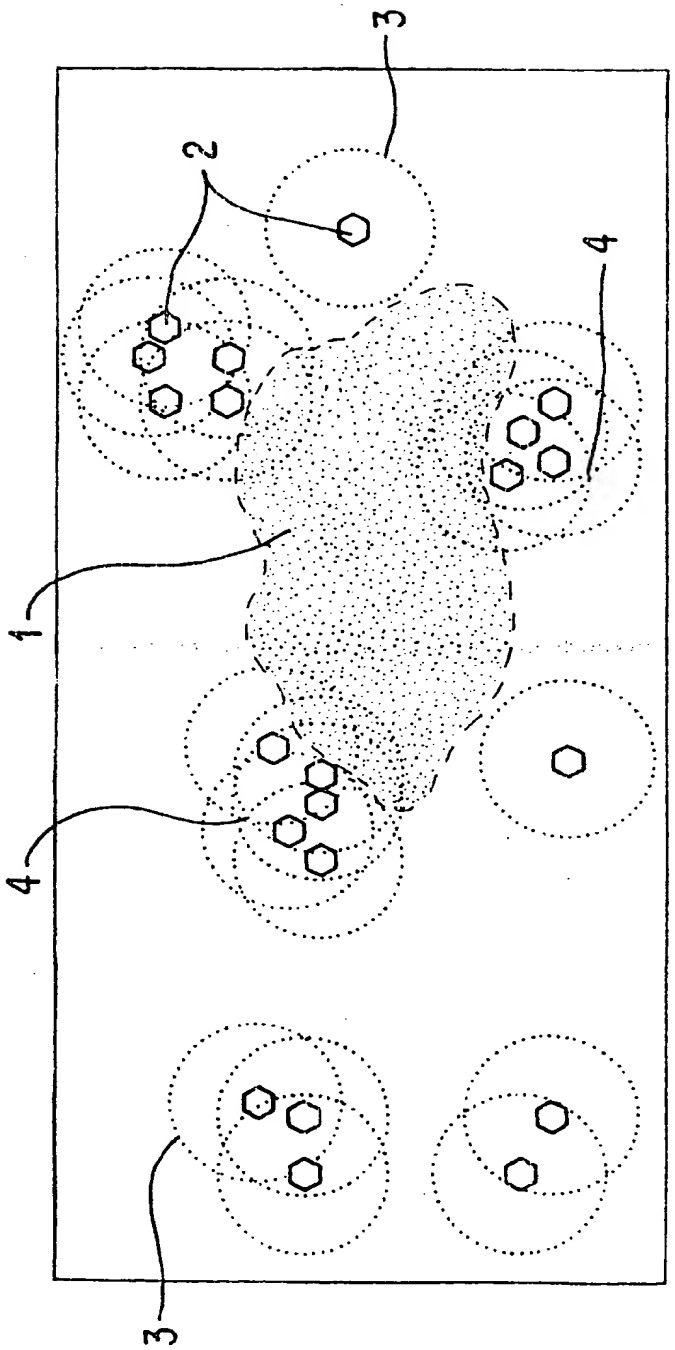


FIG. 1(a)

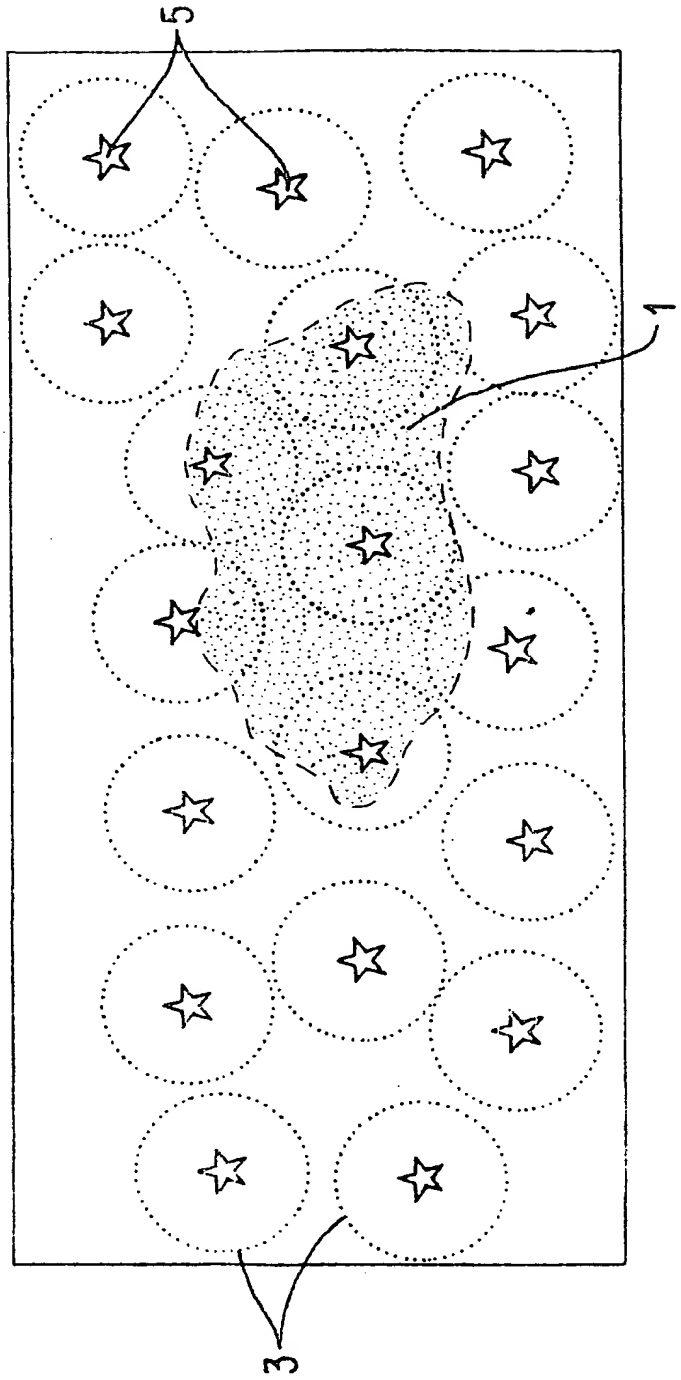


FIG.1(b)

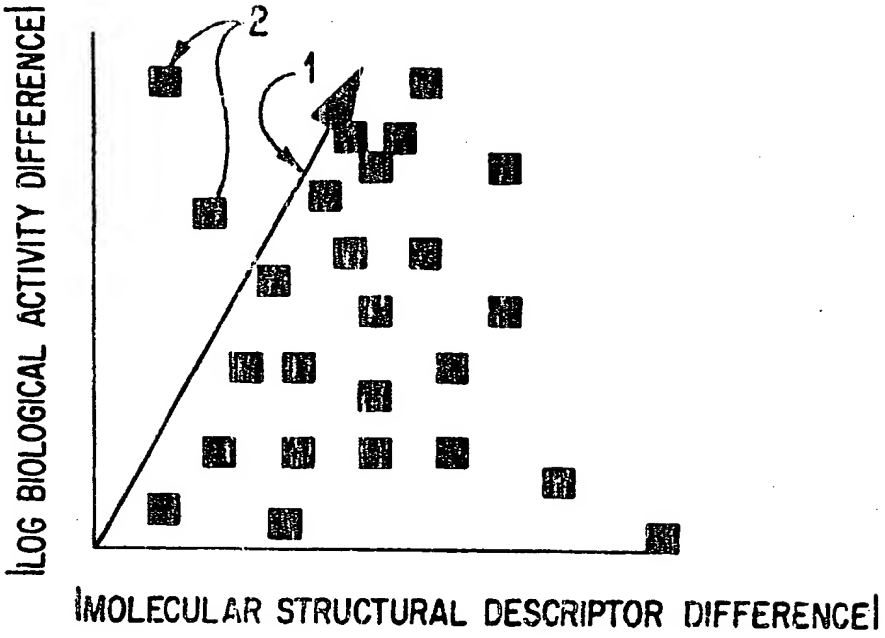
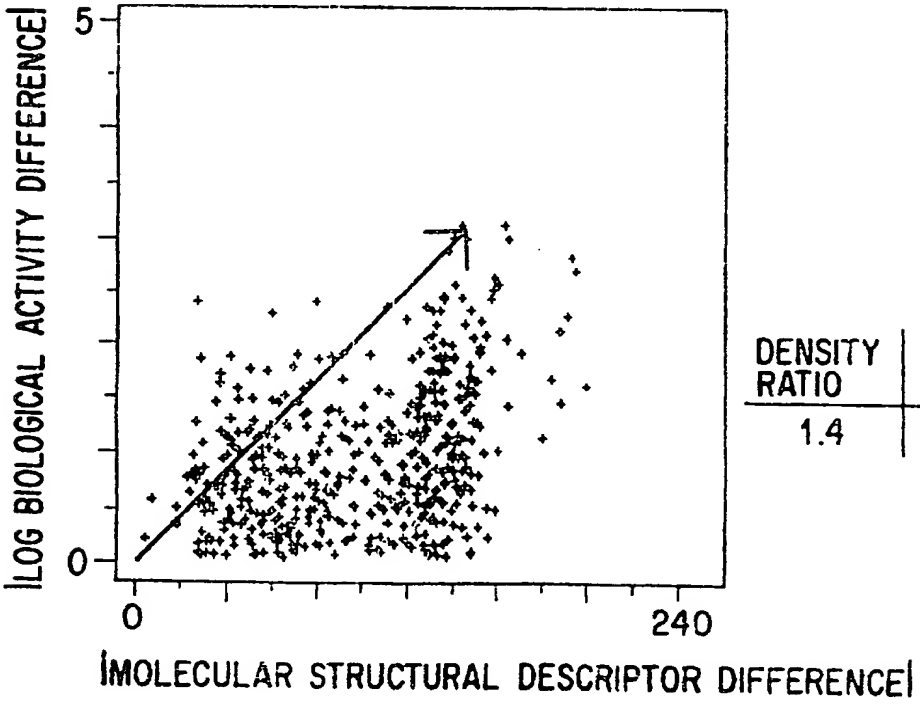
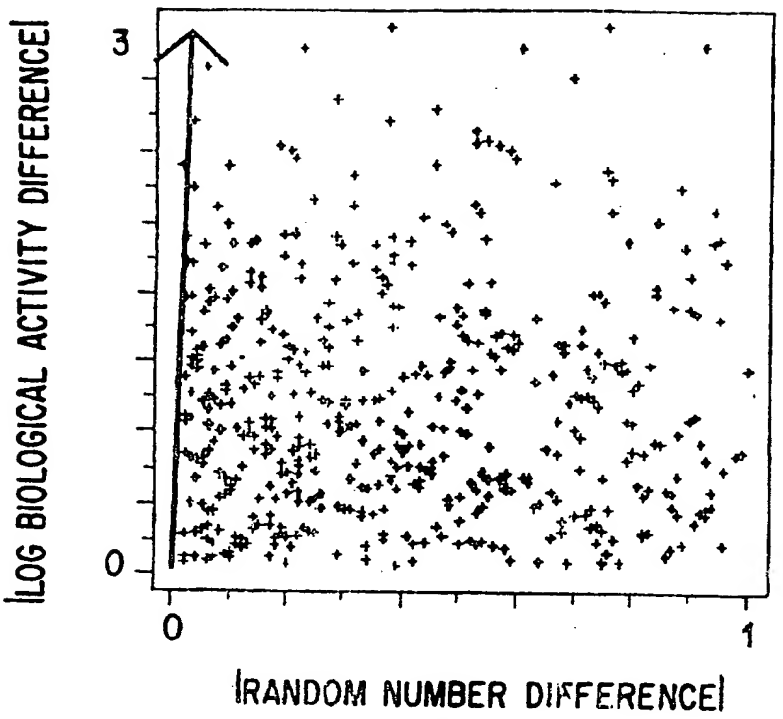


FIG. 2



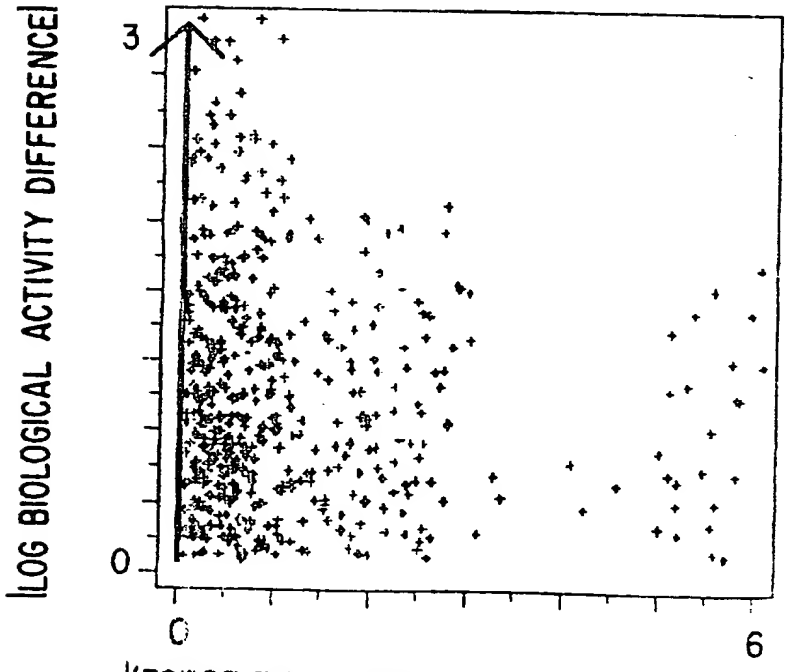
DENSITY RATIO	χ^2
1.4	51.21

FIG. 3



DENSITY RATIO	χ^2
1.00	0.00

FIG. 4



DENSITY RATIO	χ^2
0.96	0.46

FIG. 5

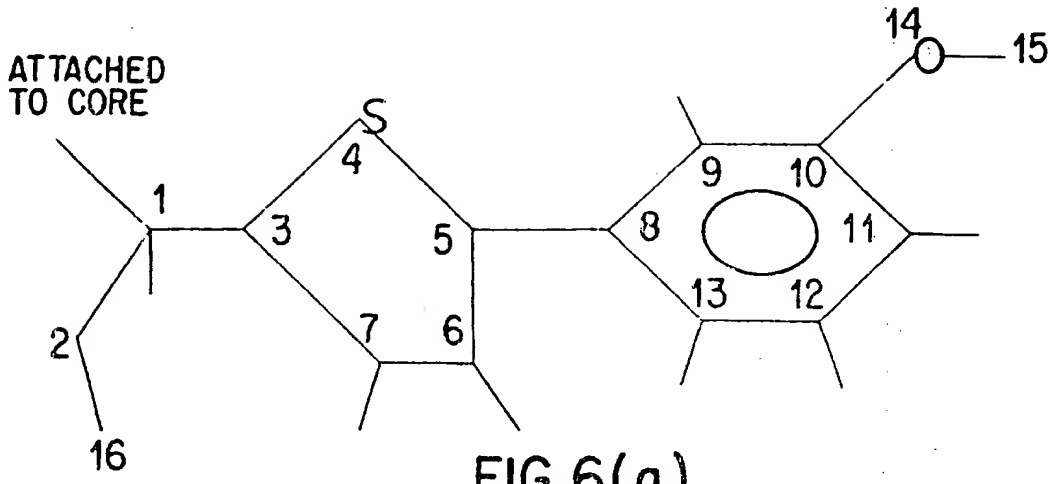


FIG. 6(a)

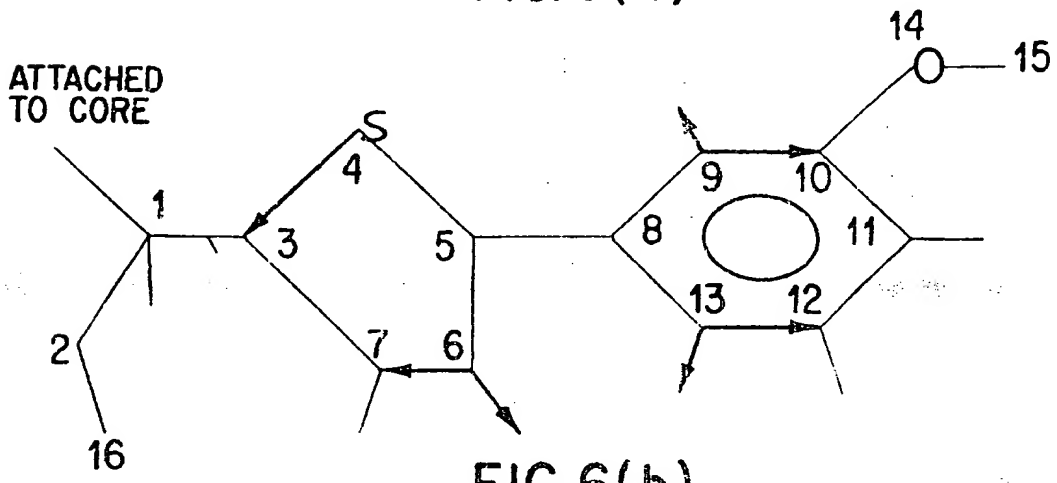


FIG. 6(b)

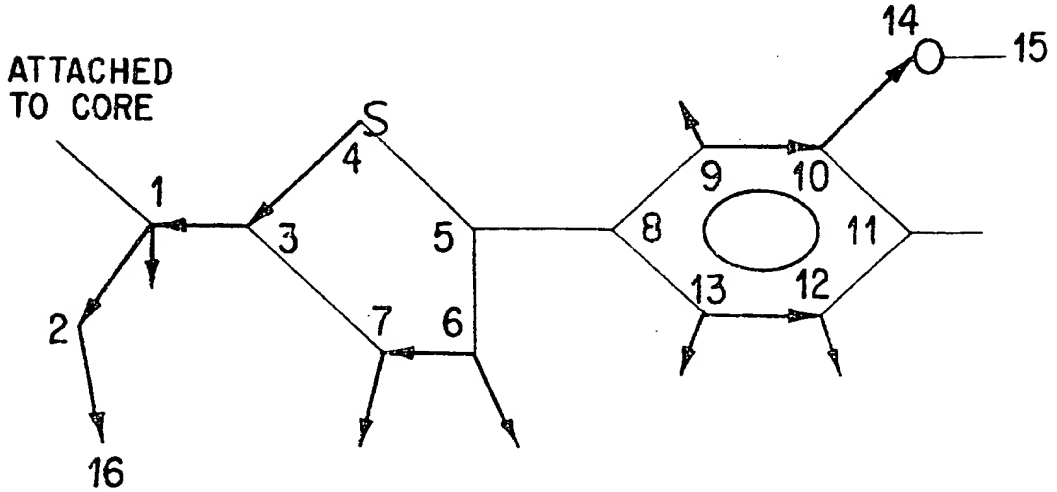


FIG. 6(c)

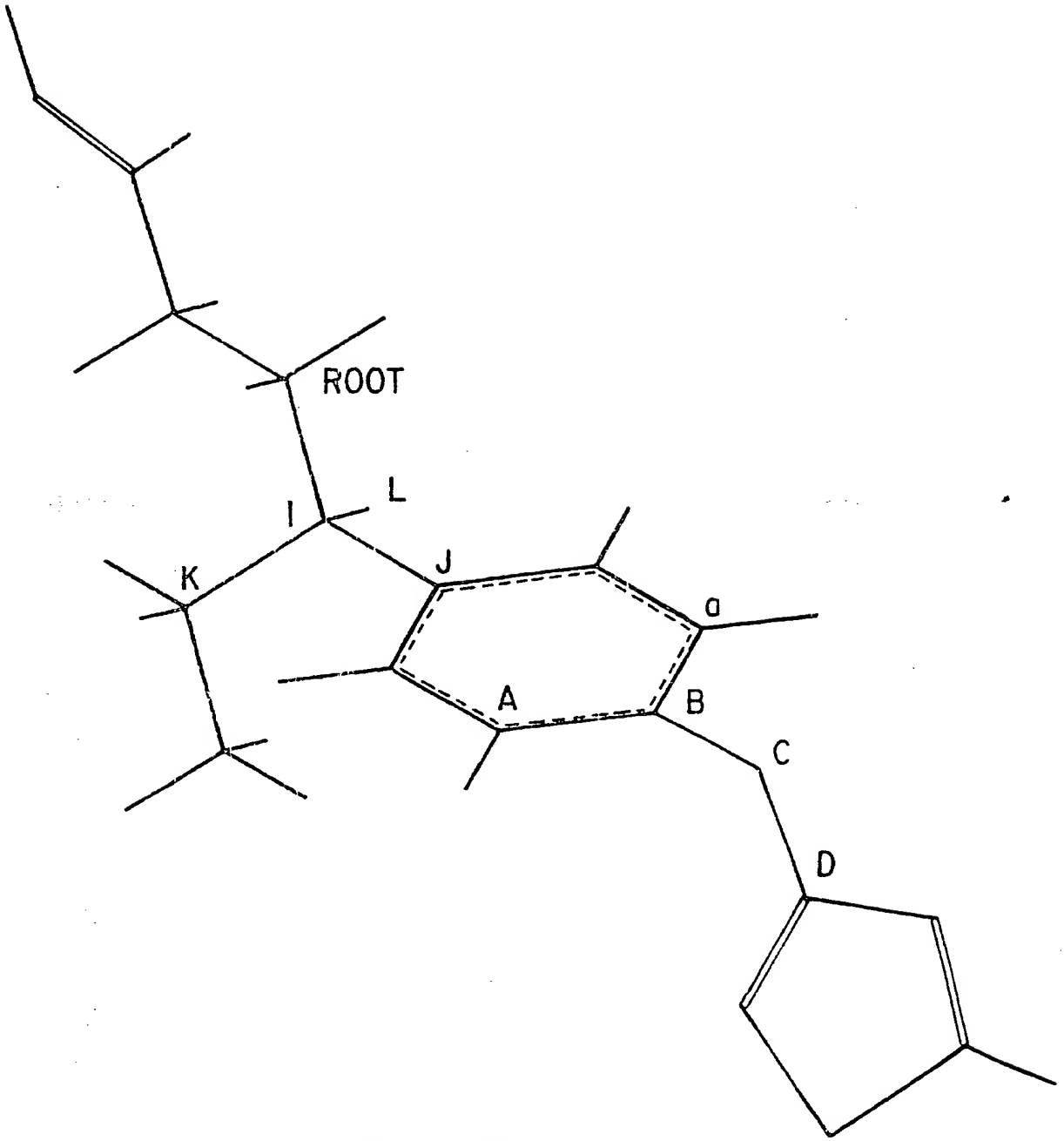


FIG. 6D

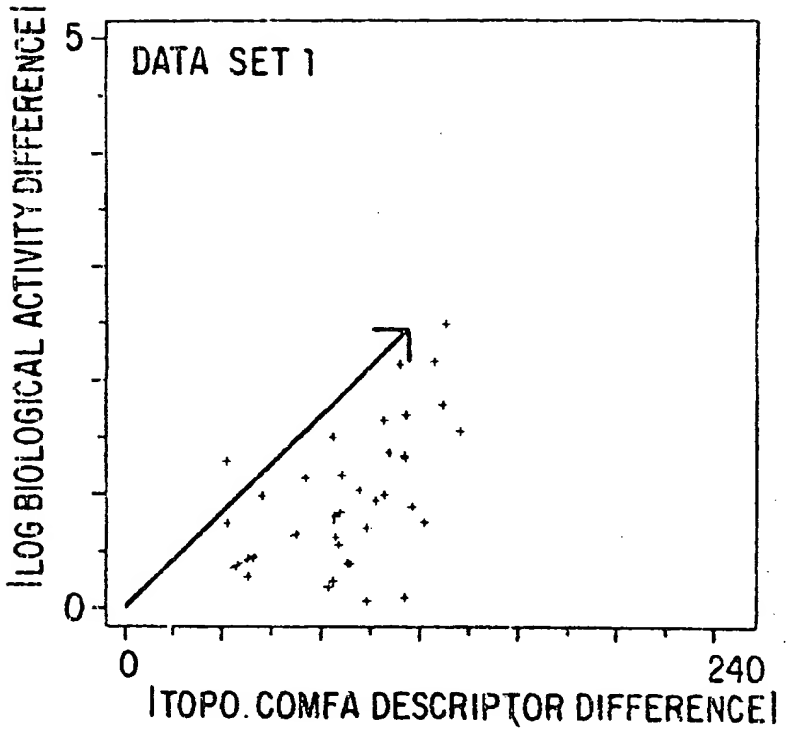

DATA SET 1
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG.7(a)

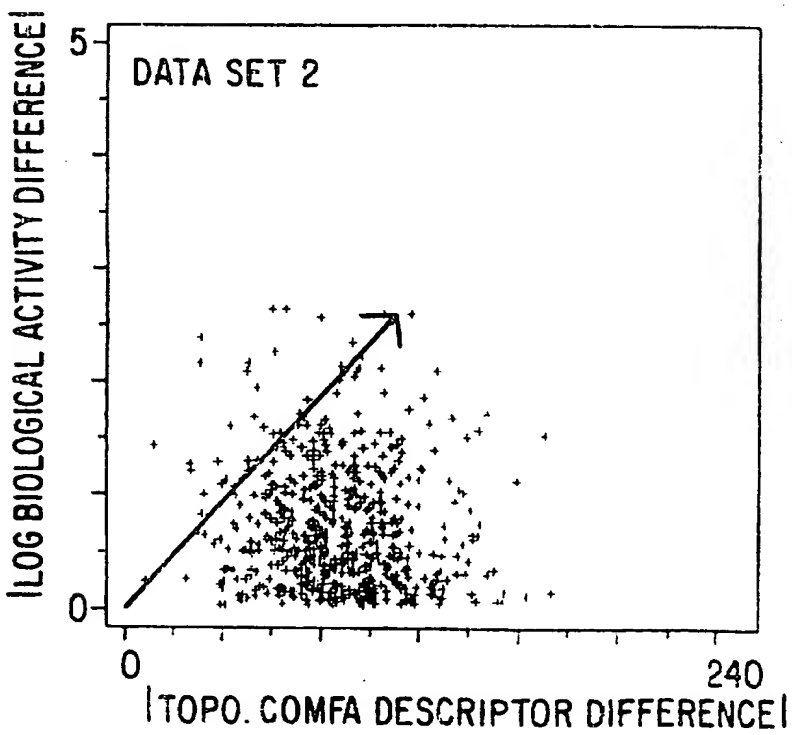

DATA SET 2
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG.7(b)

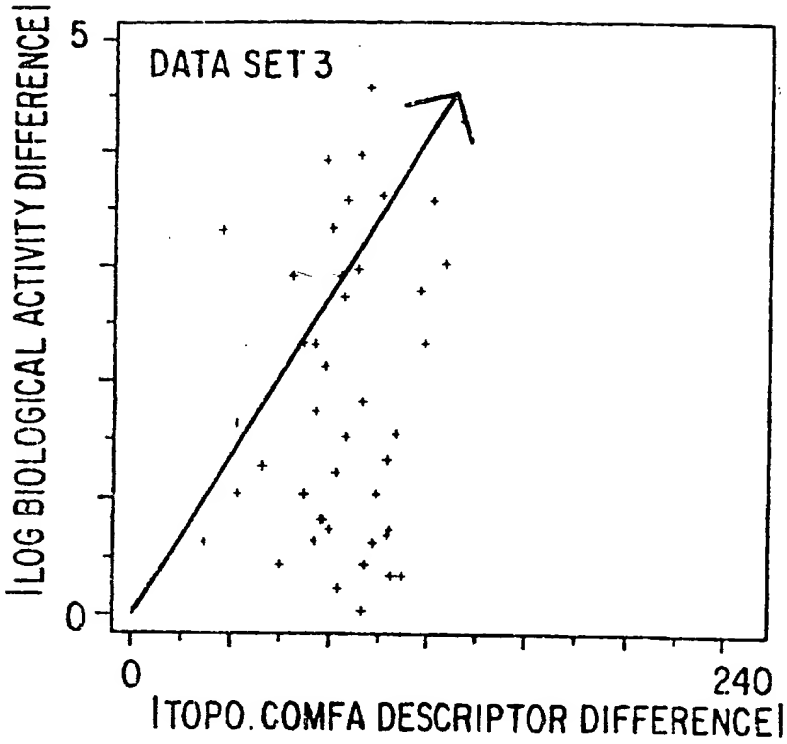

DATA SET 3
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG. 7(c)

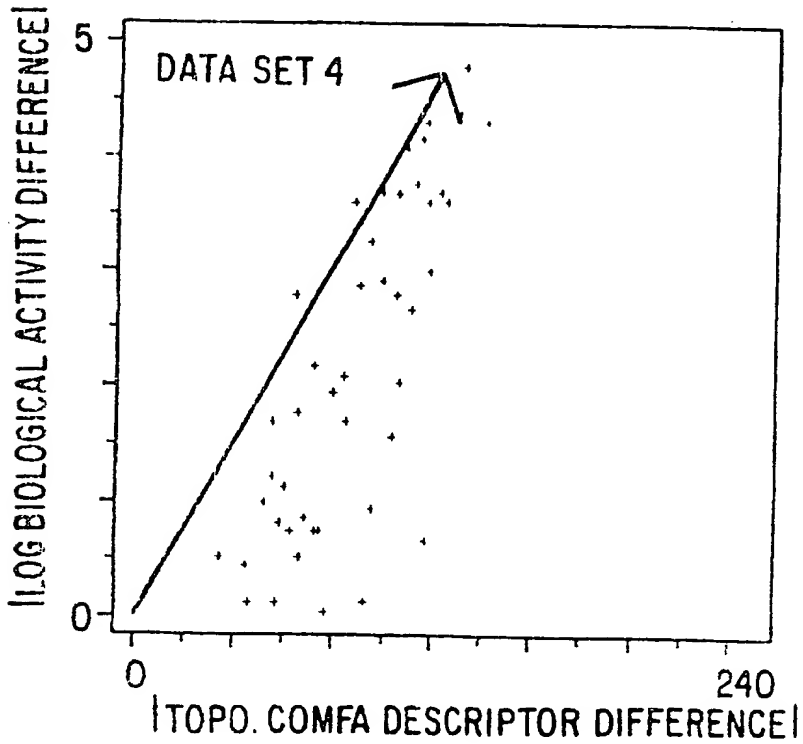

DATA SET 4
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG. 7(d)

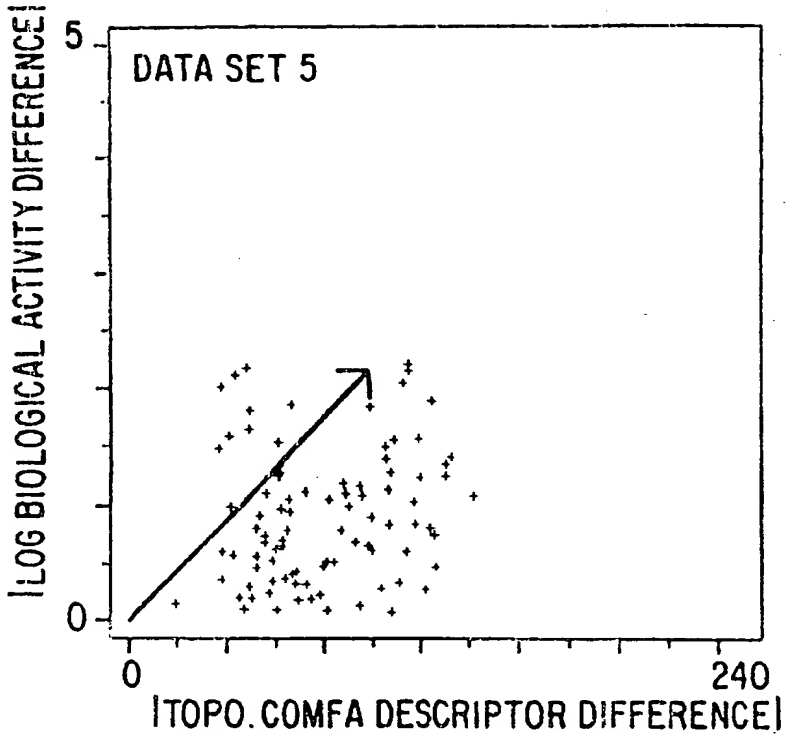

DATA SET 5
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG. 7(e)

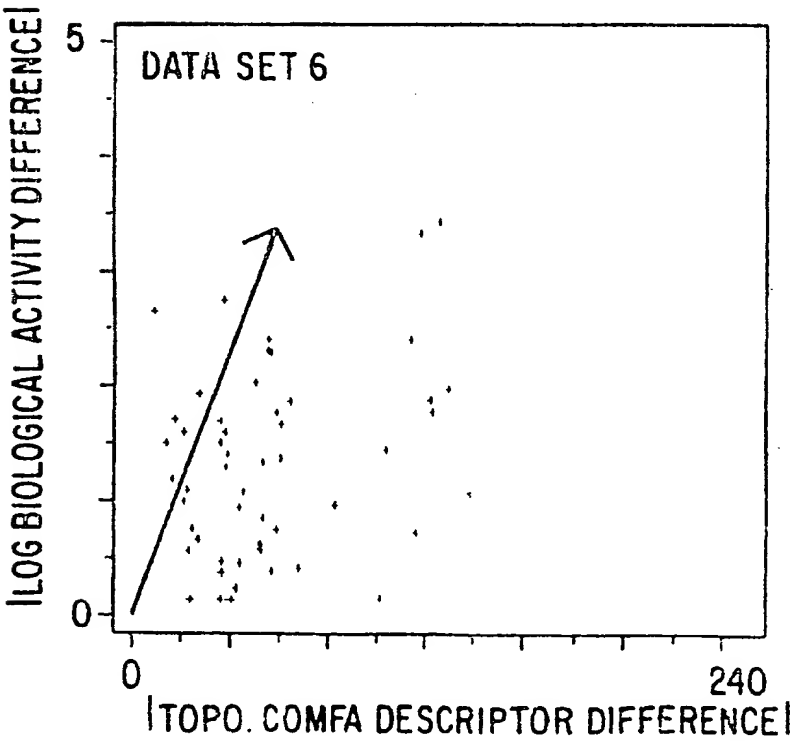

DATA SET 6
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG. 7(f)

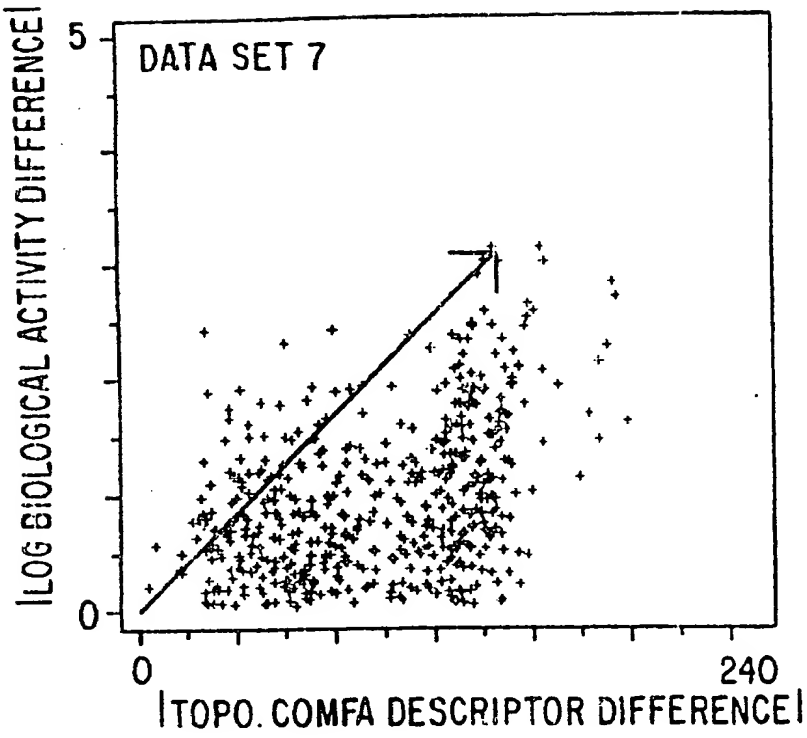

DATA SET 7
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG.7(g)

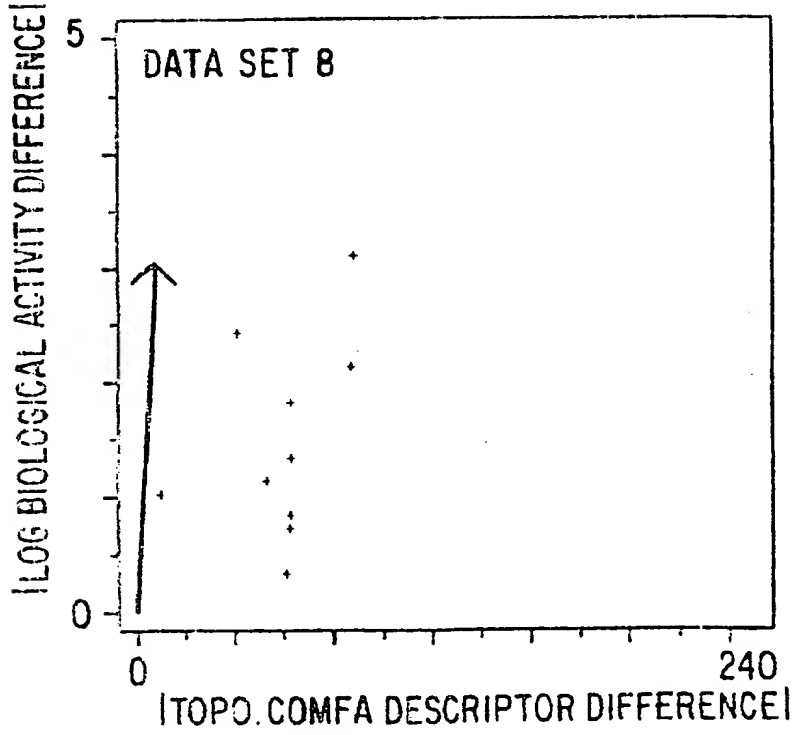

DATA SET 8
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG.7(h)

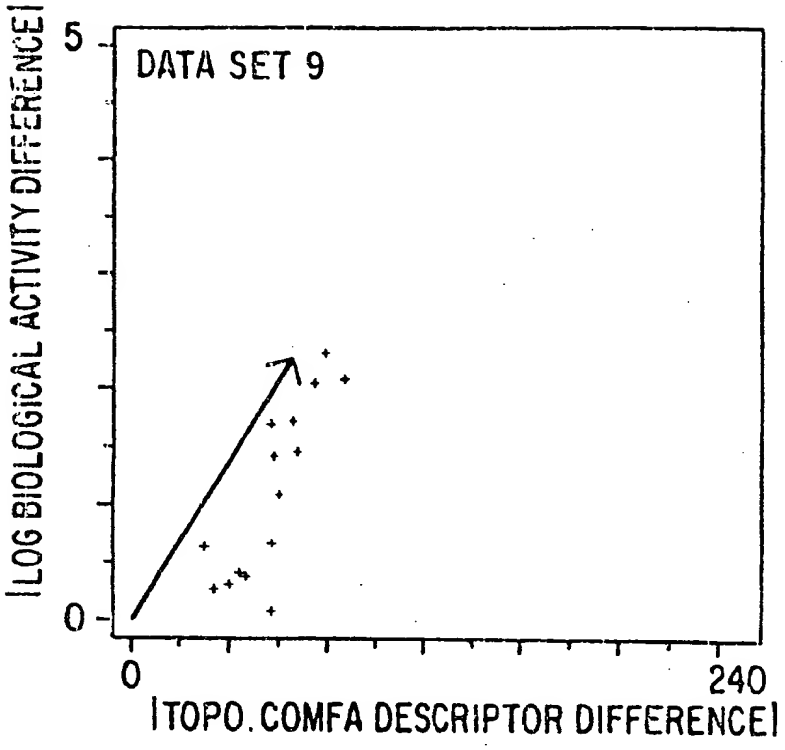

DATA SET 9
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG. 7(i)

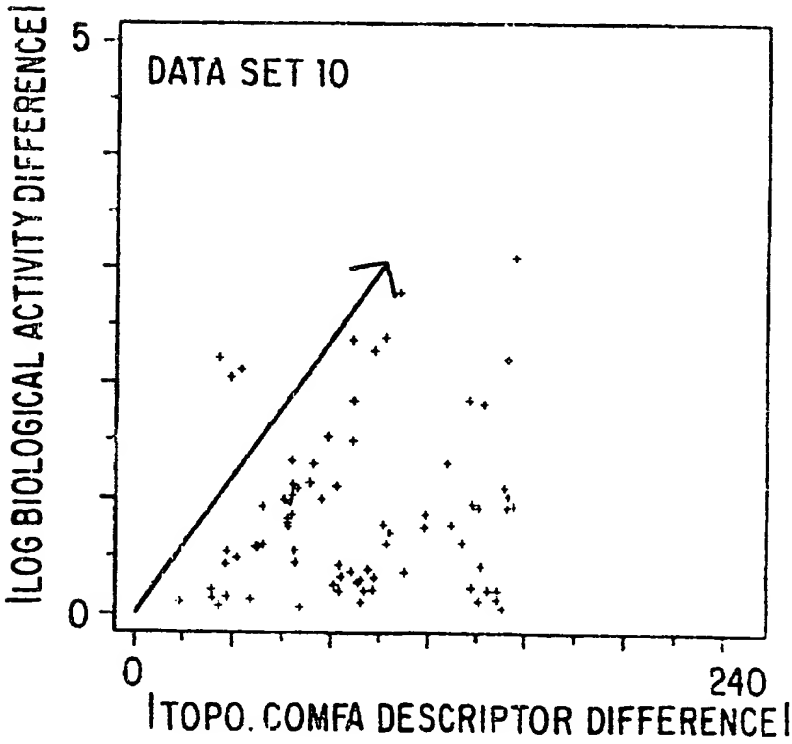

DATA SET 10
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG. 7(j)

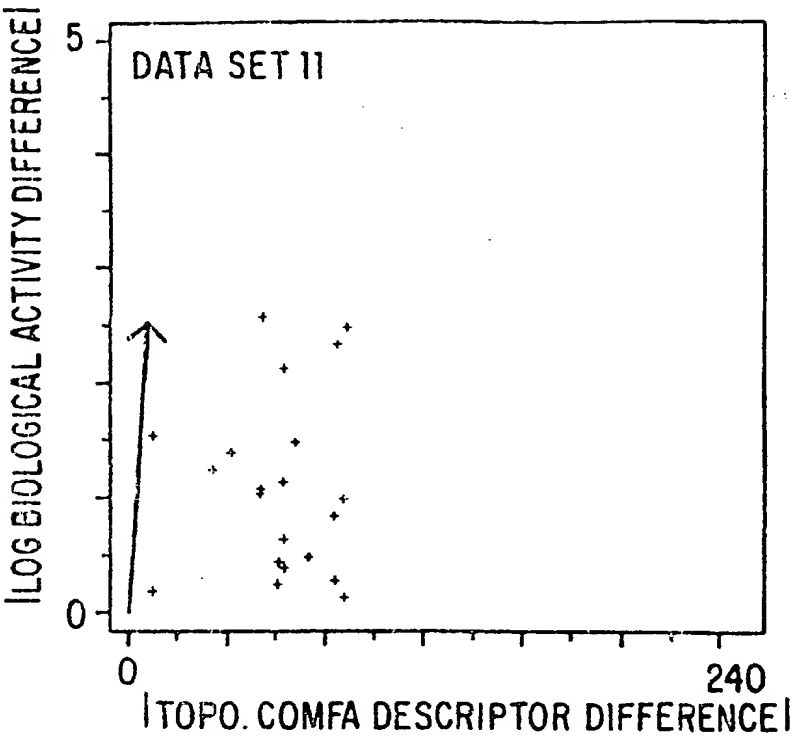

DATA SET 11
5
0
0
240
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG.7(k)

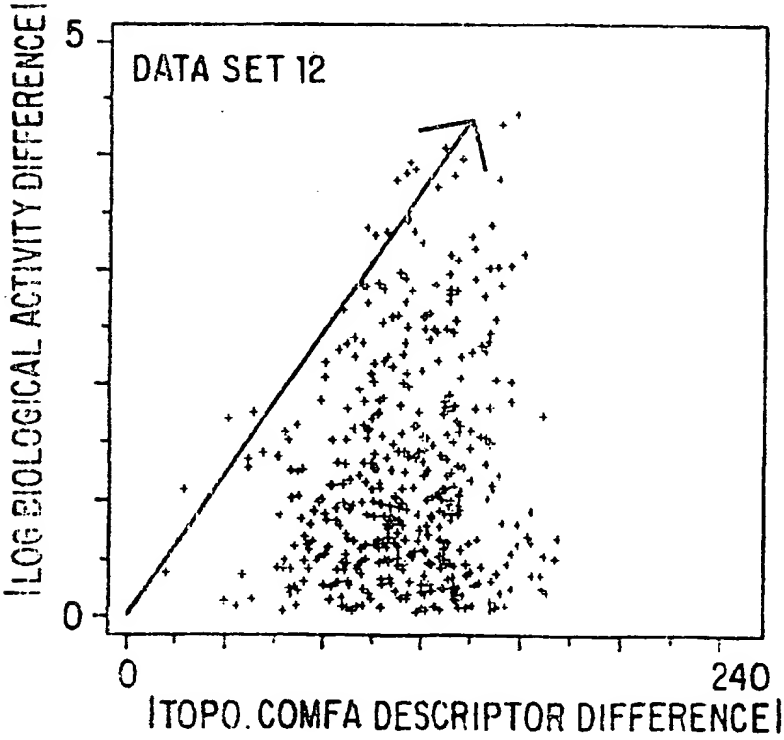

DATA SET 12
5
0
0
240
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG.7(l)

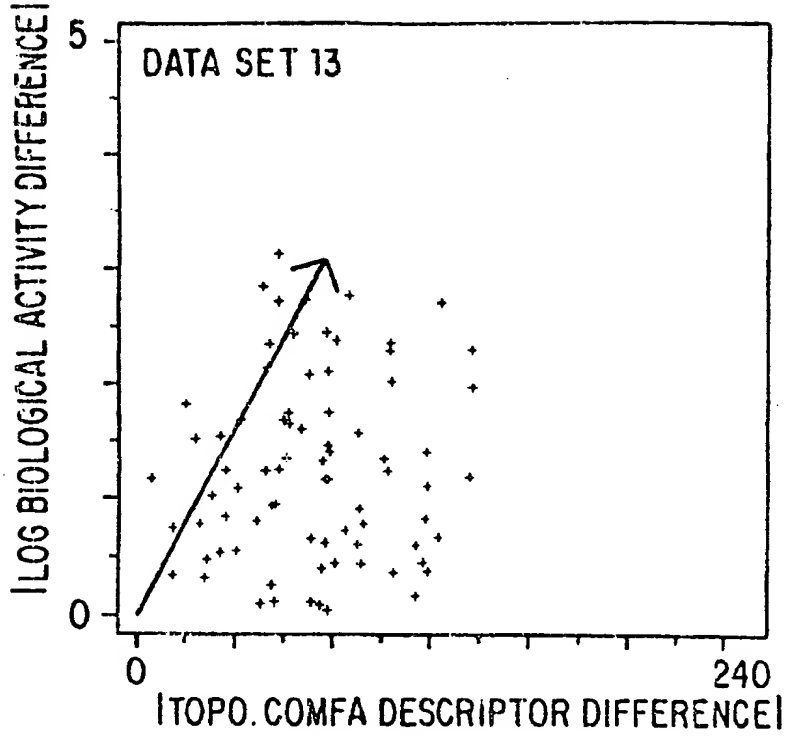

DATA SET 13
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG. 7(m)

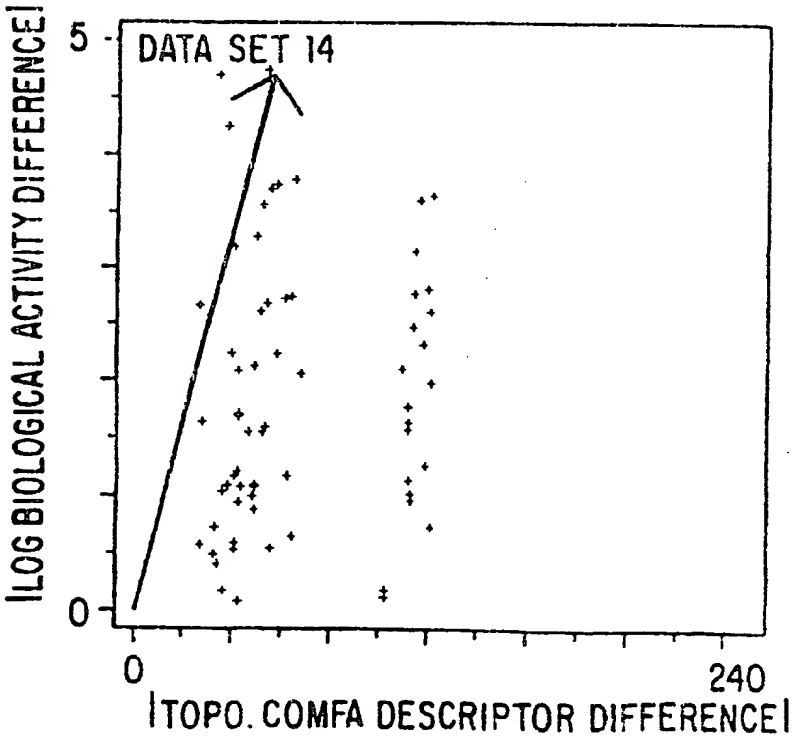

DATA SET 14
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG. 7(n)

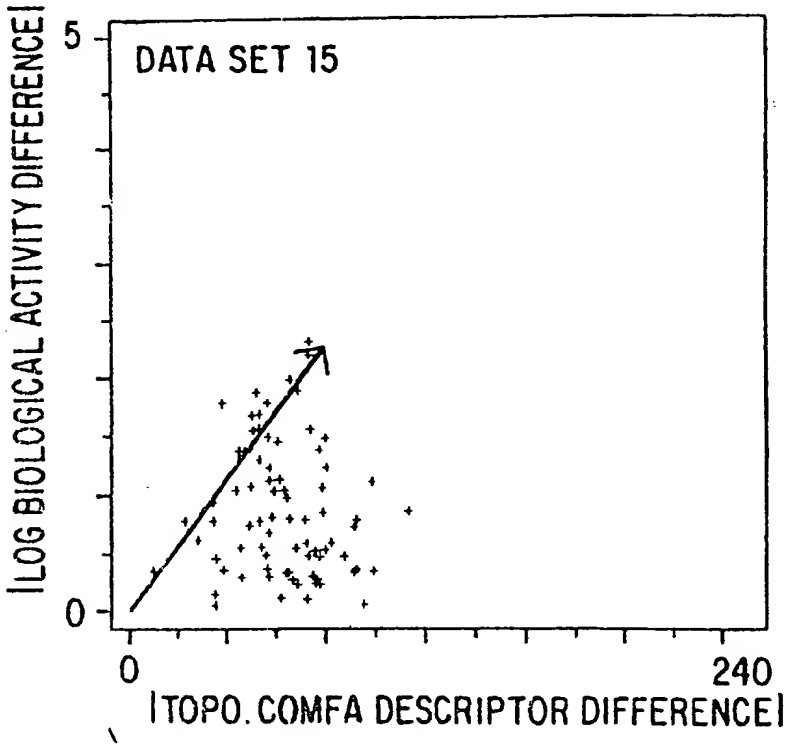

DATA SET 15
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG.7(o)

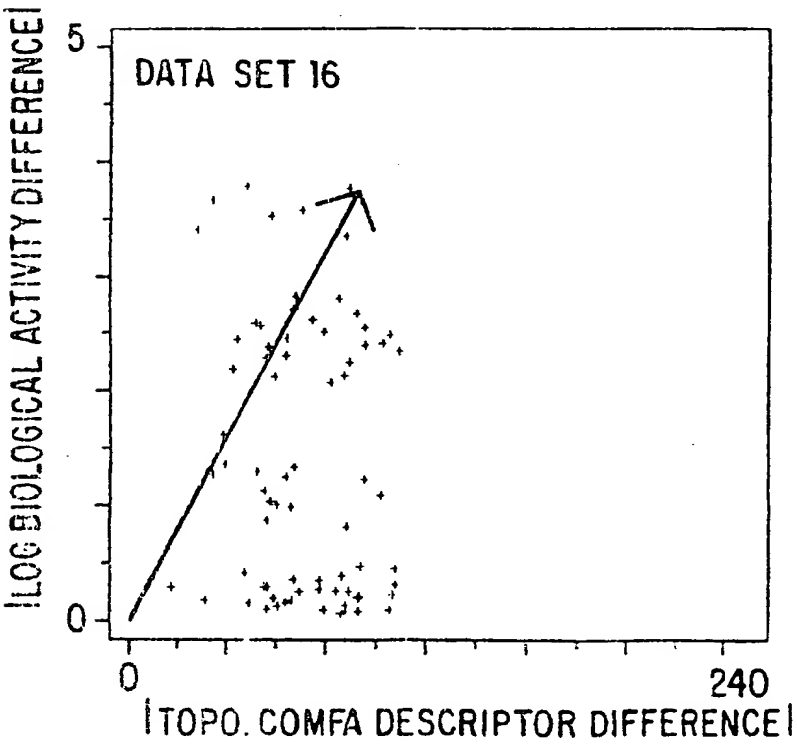

DATA SET 16
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG.7(p)

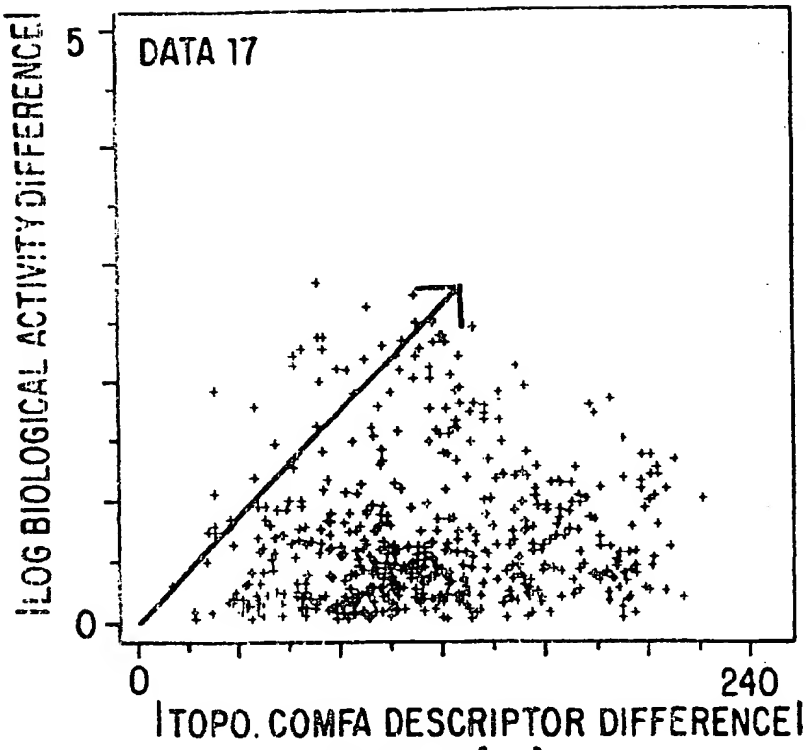

DATA 17
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG. 7(q)

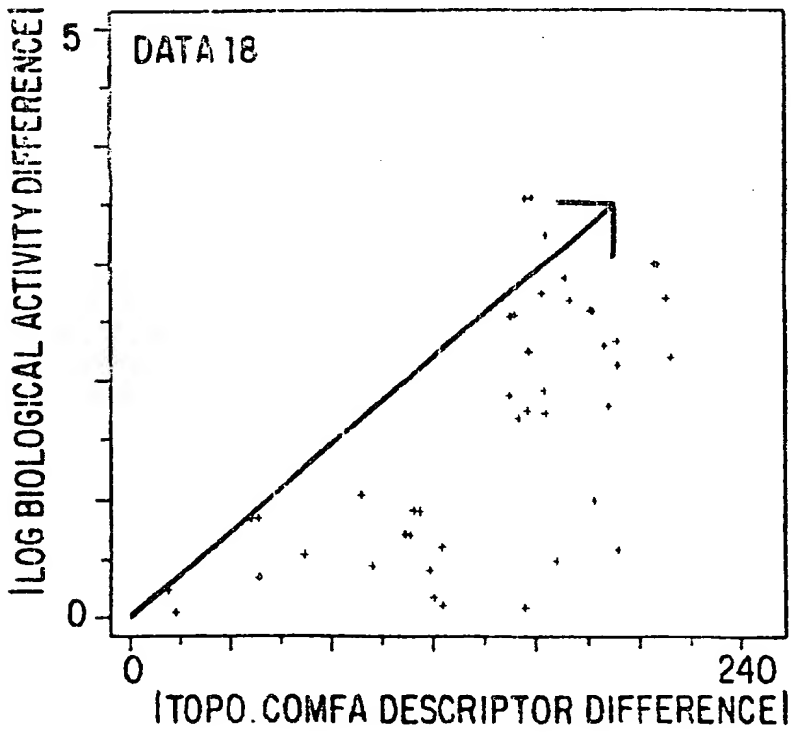

DATA 18
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
5
0
0
240
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG. 7(r)

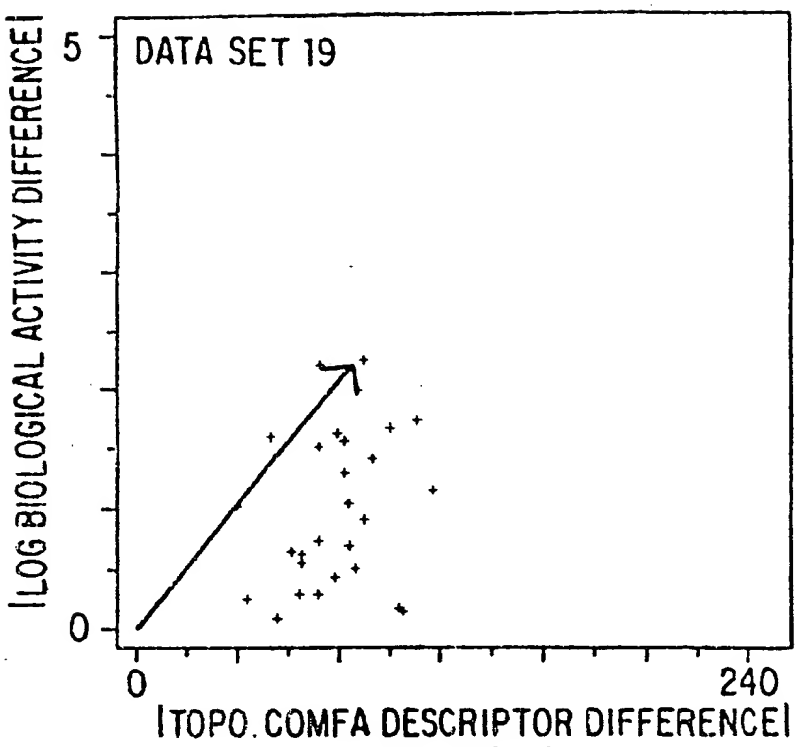

DATA SET 19
5
0
0
240
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG. 7(s)

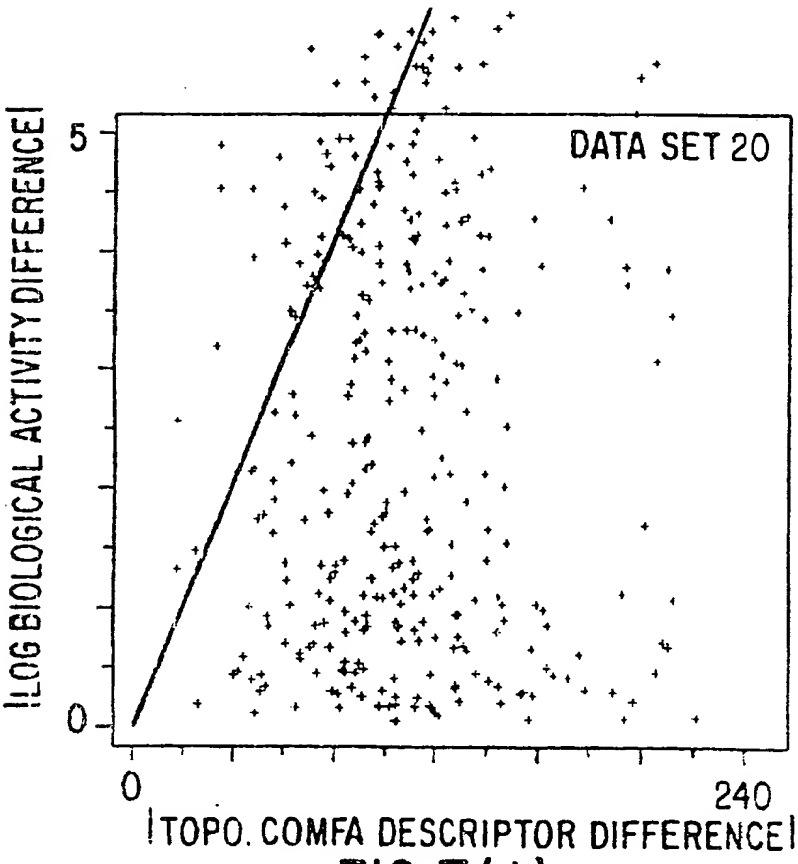

DATA SET 20
5
0
0
240
|LOG BIOLOGICAL ACTIVITY DIFFERENCE|
|TOPO. COMFA DESCRIPTOR DIFFERENCE|

FIG 7(t)

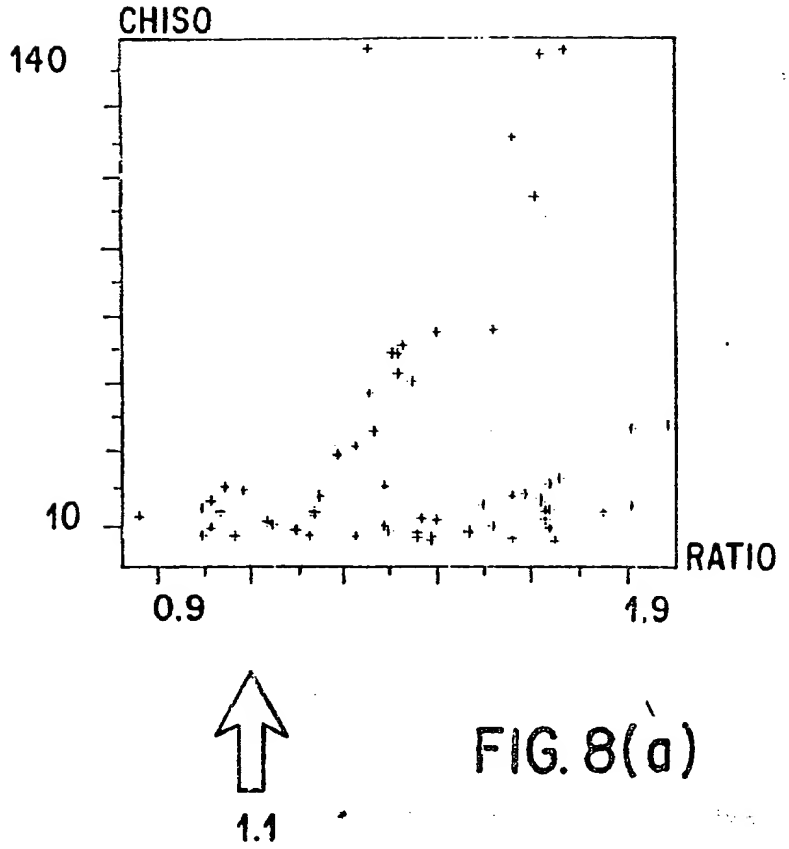


FIG. 8(a)

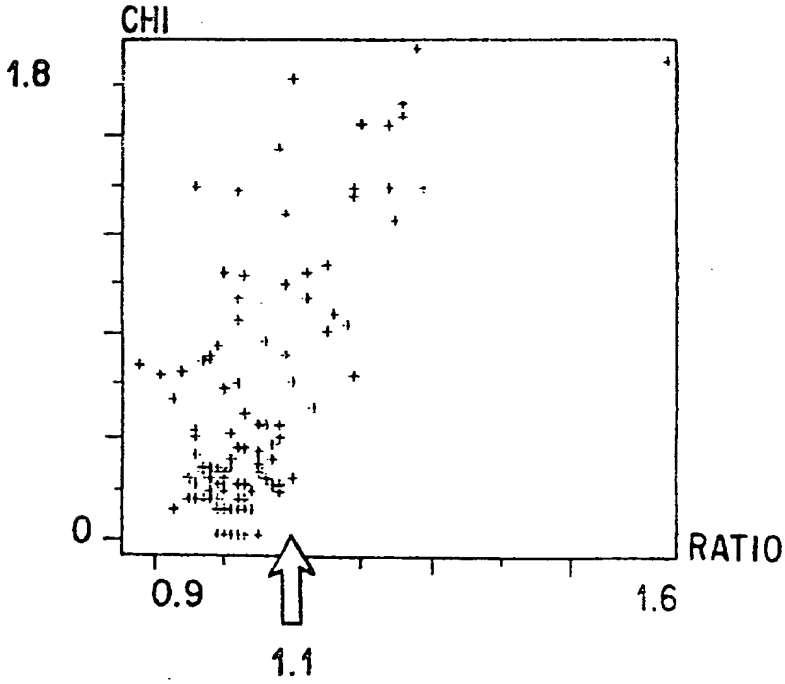


FIG. 8(b)

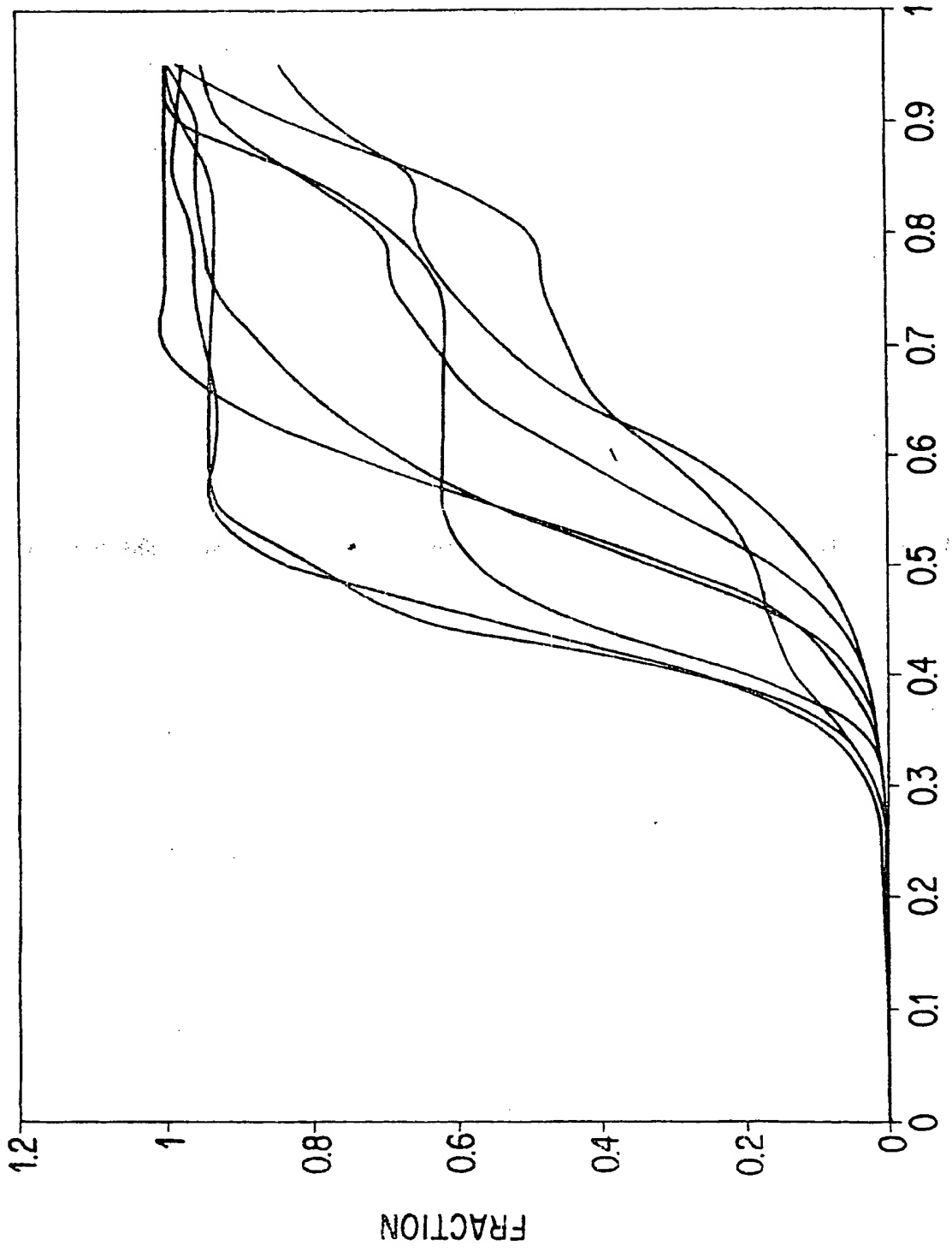


FIG. 9(a)

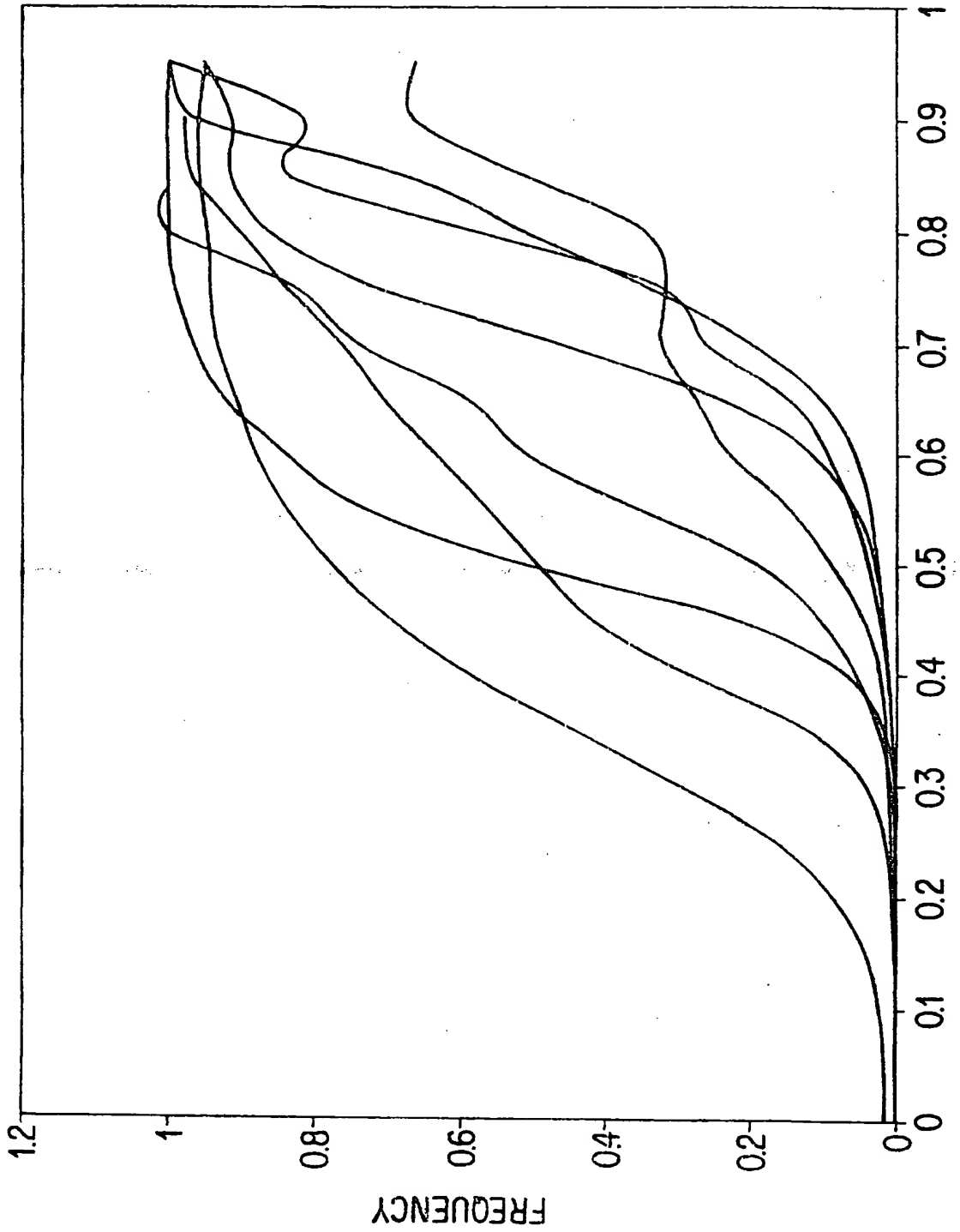


FIG.9(b)

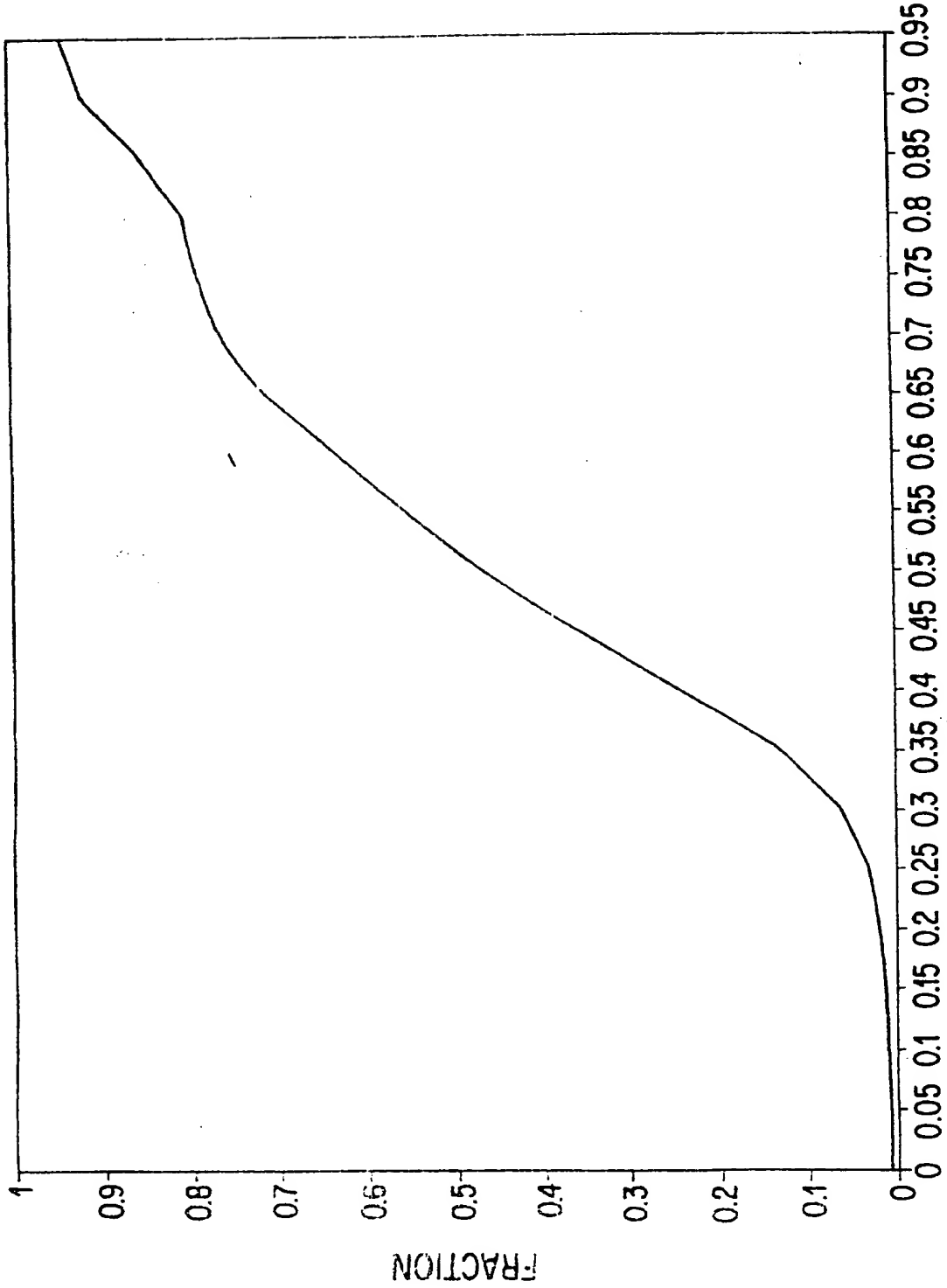

FRACTION
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95
TANIMOTO

FIG. 9(c)

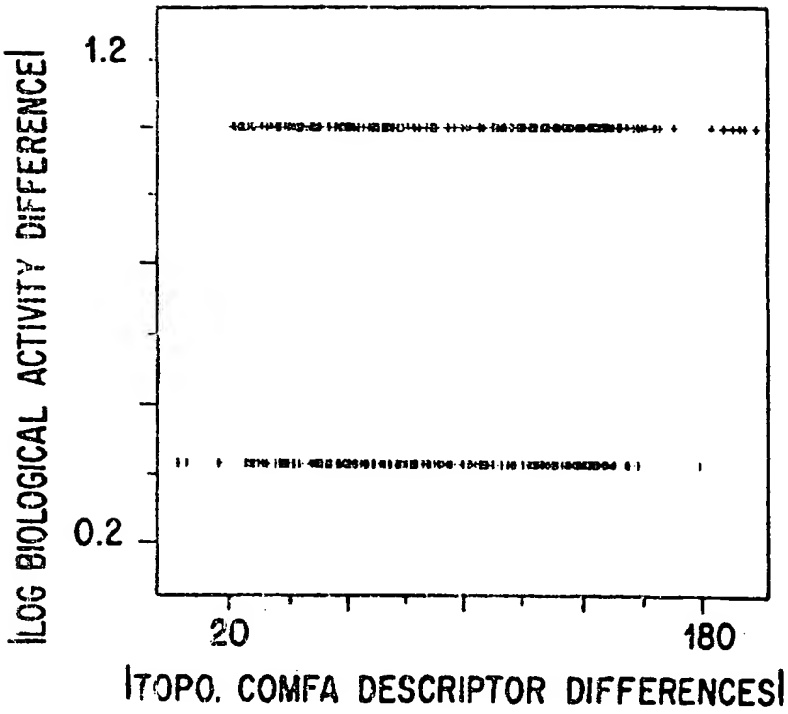


FIG. 10(a)

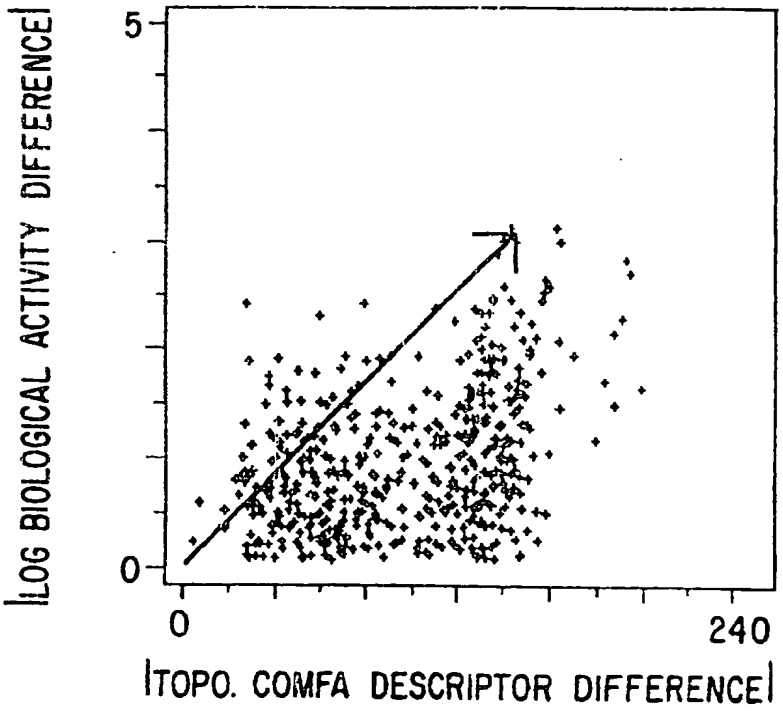


FIG. 10(b)

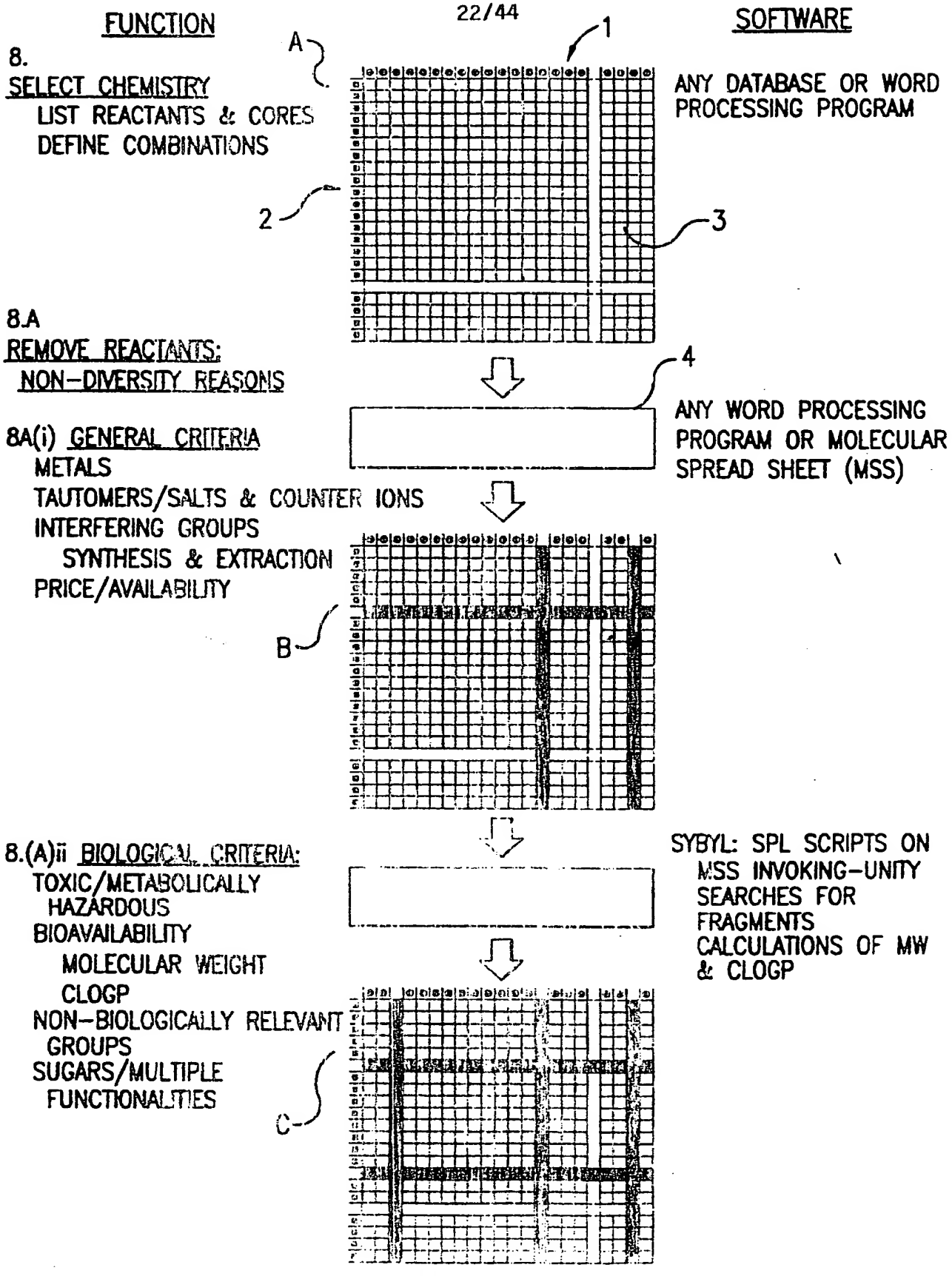

FUNCTION
SOFTWARE
8.
SELECT CHEMISTRY
LIST REACTANTS & CORES
DEFINE COMBINATIONS
A
1
2
3
ANY DATABASE OR WORD PROCESSING PROGRAM
8.A
REMOVE REACTANTS:
NON-DIVERSITY REASONS
4
8A(i) GENERAL CRITERIA
METALS
TAUTOMERS/SALTS & COUNTER IONS
INTERFERING GROUPS
SYNTHESIS & EXTRACTION
PRICE/AVAILABILITY
ANY WORD PROCESSING PROGRAM OR MOLECULAR SPREAD SHEET (MSS)
B
8.(A)ii BIOLOGICAL CRITERIA:
TOXIC/METABOLICALLY HAZARDOUS
BIOAVAILABILITY
MOLECULAR WEIGHT
CLOGP
NON-BIOLOGICALLY RELEVANT GROUPS
SUGARS/MULTIPLE FUNCTIONALITIES
SYBYL: SPL SCRIPTS ON MSS INVOKING-UNITY SEARCHES FOR FRAGMENTS CALCULATIONS OF MW & CLOGP
C

FIG. 1(a)

8B

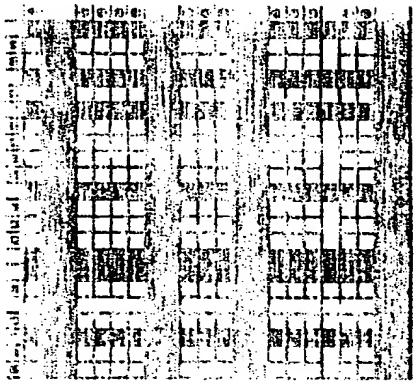
REMOVE NON-DIVERSE REACTANTS

3D STRUCTURE GENERATION
TOPOMERIC CONFORMER ALIGNMENT
CoMFA FIELD GENERATION:
HYDROGEN BOND FIELD GENERATION
ROTATABLE BOND FIELD ATTENUATION
CALCULATE FIELD DIFFERENCES
CLUSTER
REACTANT SELECTION FROM EACH CLUSTER

SYBYL: SELECTOR & CONCORD 3.2.1 SPECIALIZED SOFTWARE: APPENDIX "A" (SYBYL: CoMFA STERIC & MSS)

SYBYL: HIERCHICAL CLUSTER – MSS
SYLBL: MSS

D

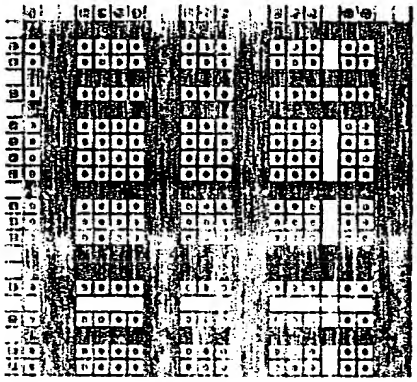


8C

COMBINE REACTANTS TO BUILD PRODUCTS

SYBYL: LEGION MSS

E



8D

REMOVE PRODUCTS:
NON-DIVERSITY REASONS

GENERAL CRITERIA:

SYBYL: SPL SCRIPTS ON MSS

F

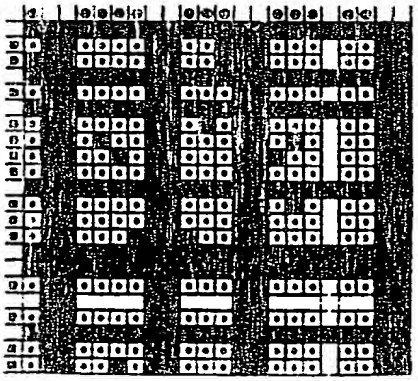


FIG. 11(b)

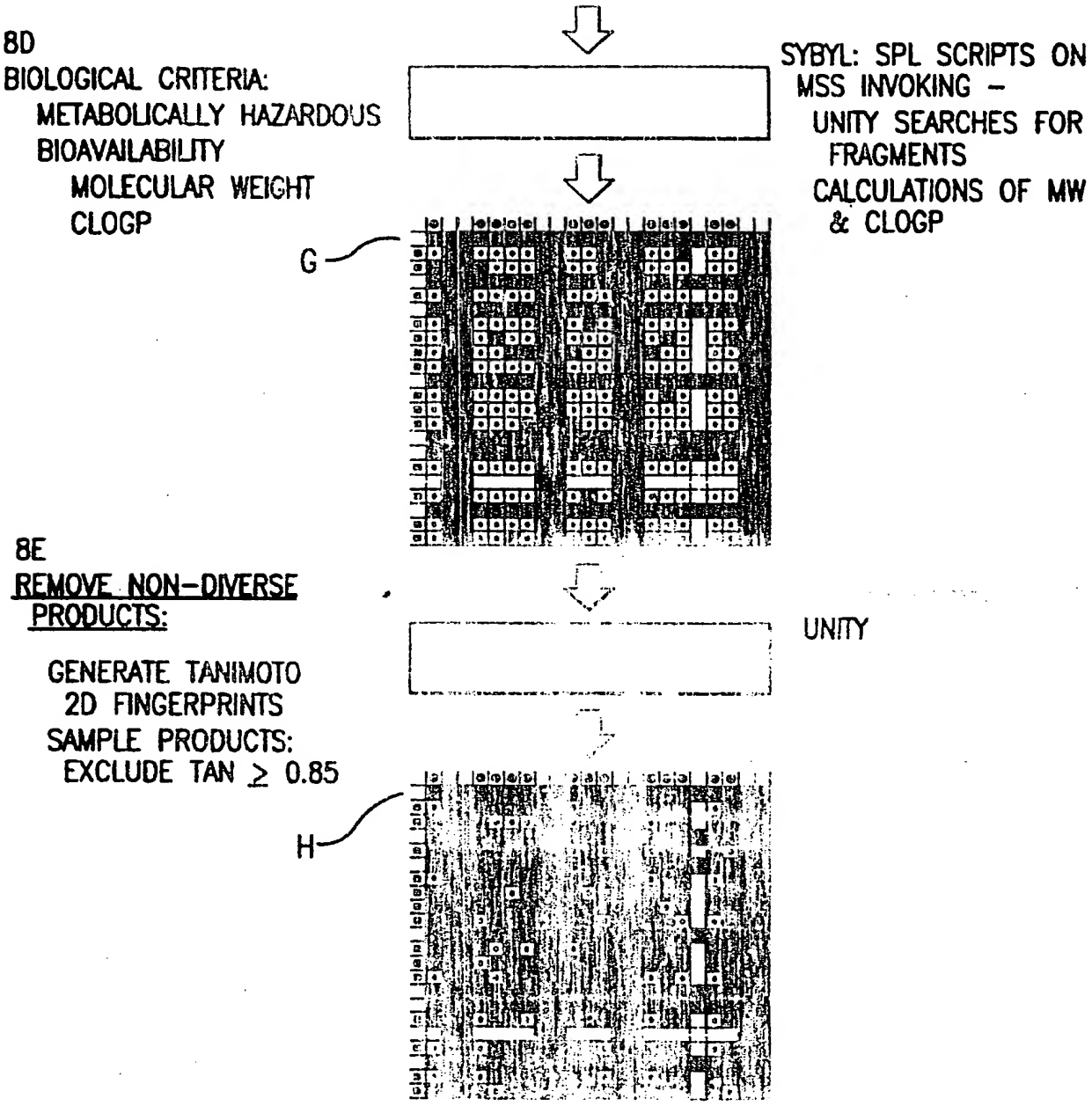
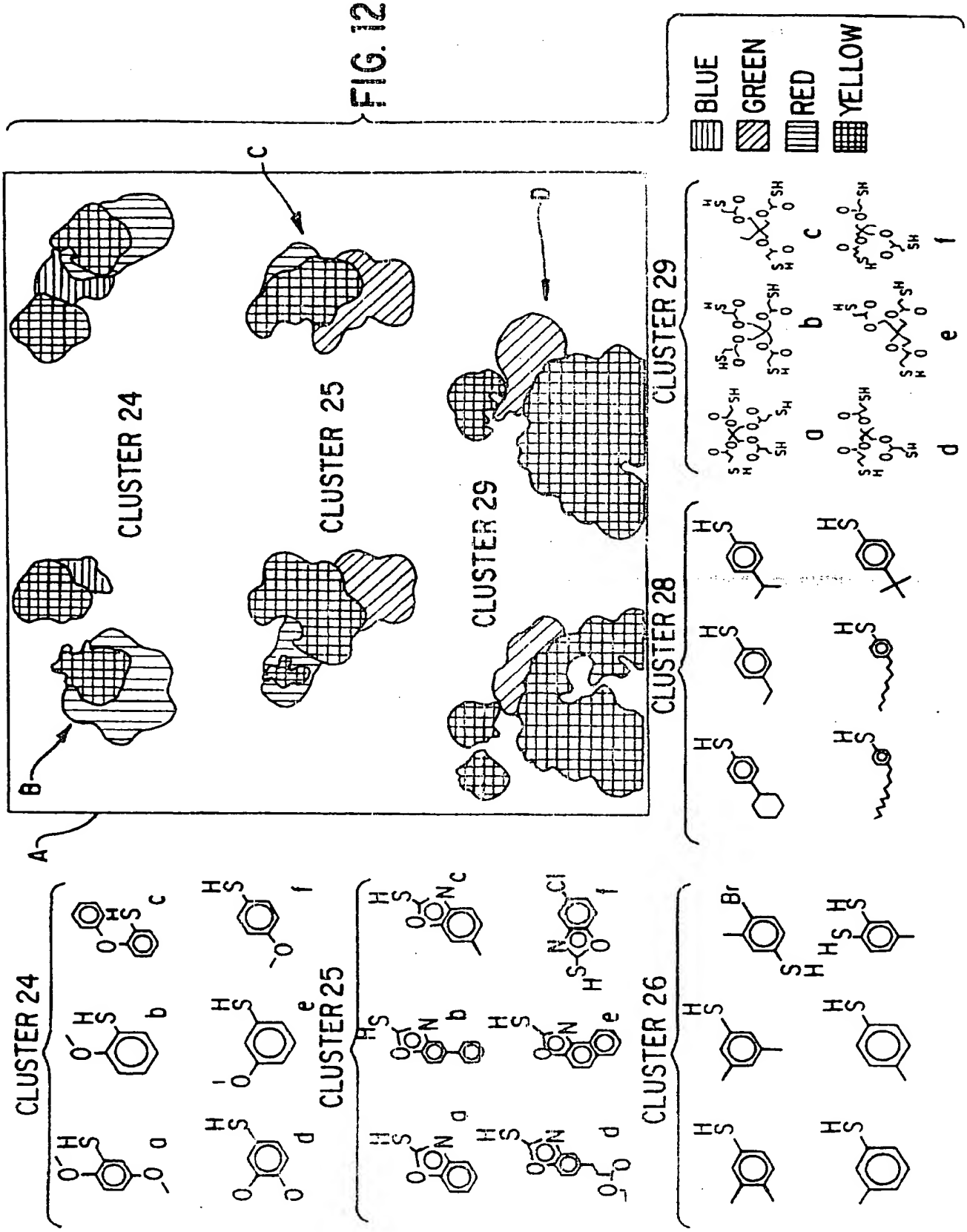

8D
BIOLOGICAL CRITERIA:
METABOLICALLY HAZARDOUS
BIOAVAILABILITY
MOLECULAR WEIGHT
CLOGP
SYBYL: SPL SCRIPTS ON MSS INVOKING –
UNITY SEARCHES FOR FRAGMENTS
CALCULATIONS OF MW & CLOGP
G
8E
REMOVE NON-DIVERSE PRODUCTS:
GENERATE TANIMOTO 2D FINGERPRINTS
SAMPLE PRODUCTS:
EXCLUDE TAN ≥ 0.85
UNITY
H

FIG. 12


BLUE
GREEN
RED
YELLOW
A
B
C
D
CLUSTER 24
CLUSTER 25
CLUSTER 29
CLUSTER 29
CLUSTER 28
CLUSTER 24
CLUSTER 25
CLUSTER 26

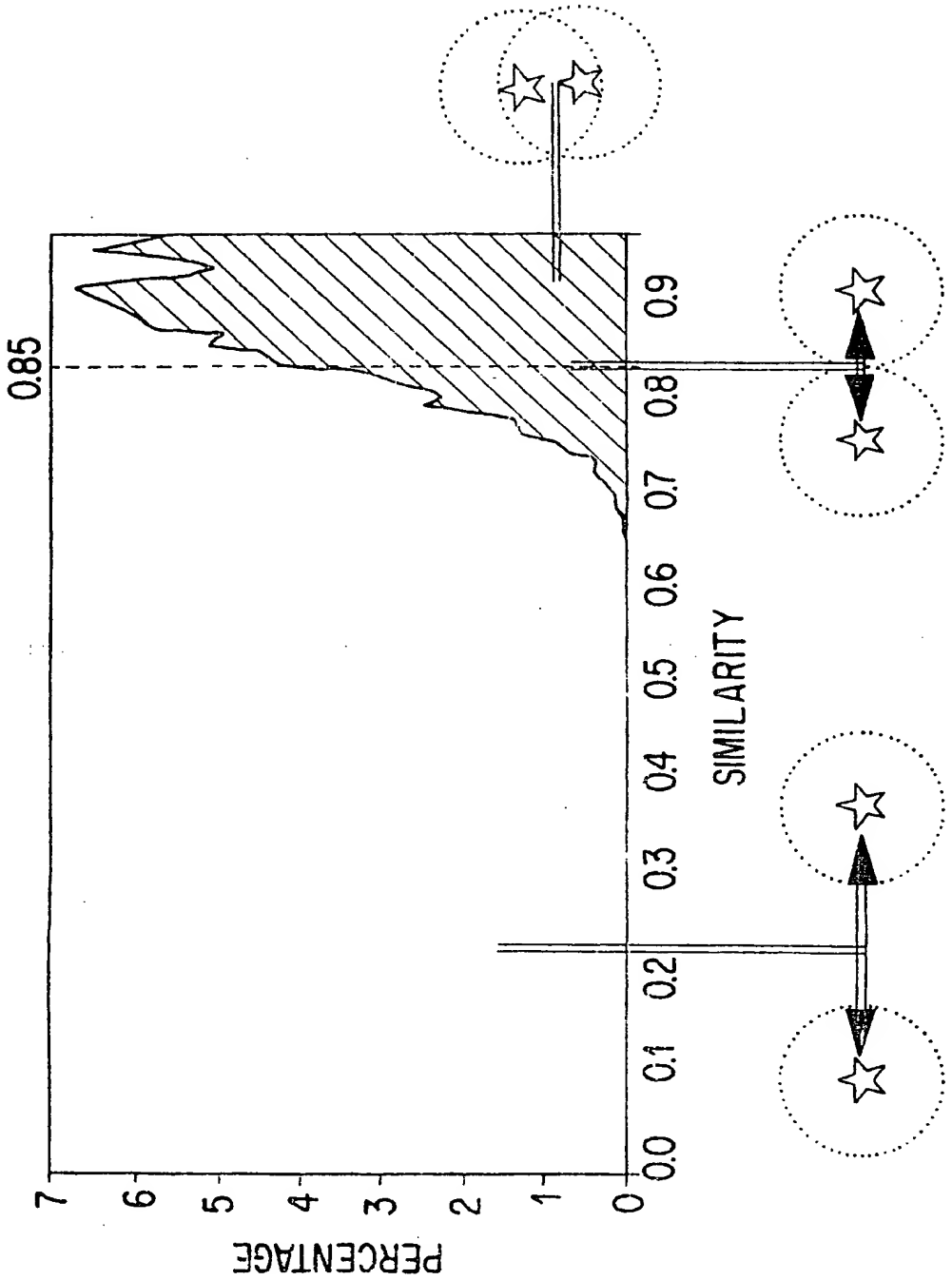

PERCENTAGE
0
1
2
3
4
5
6
7
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
SIMILARITY
0.85

FIG.13

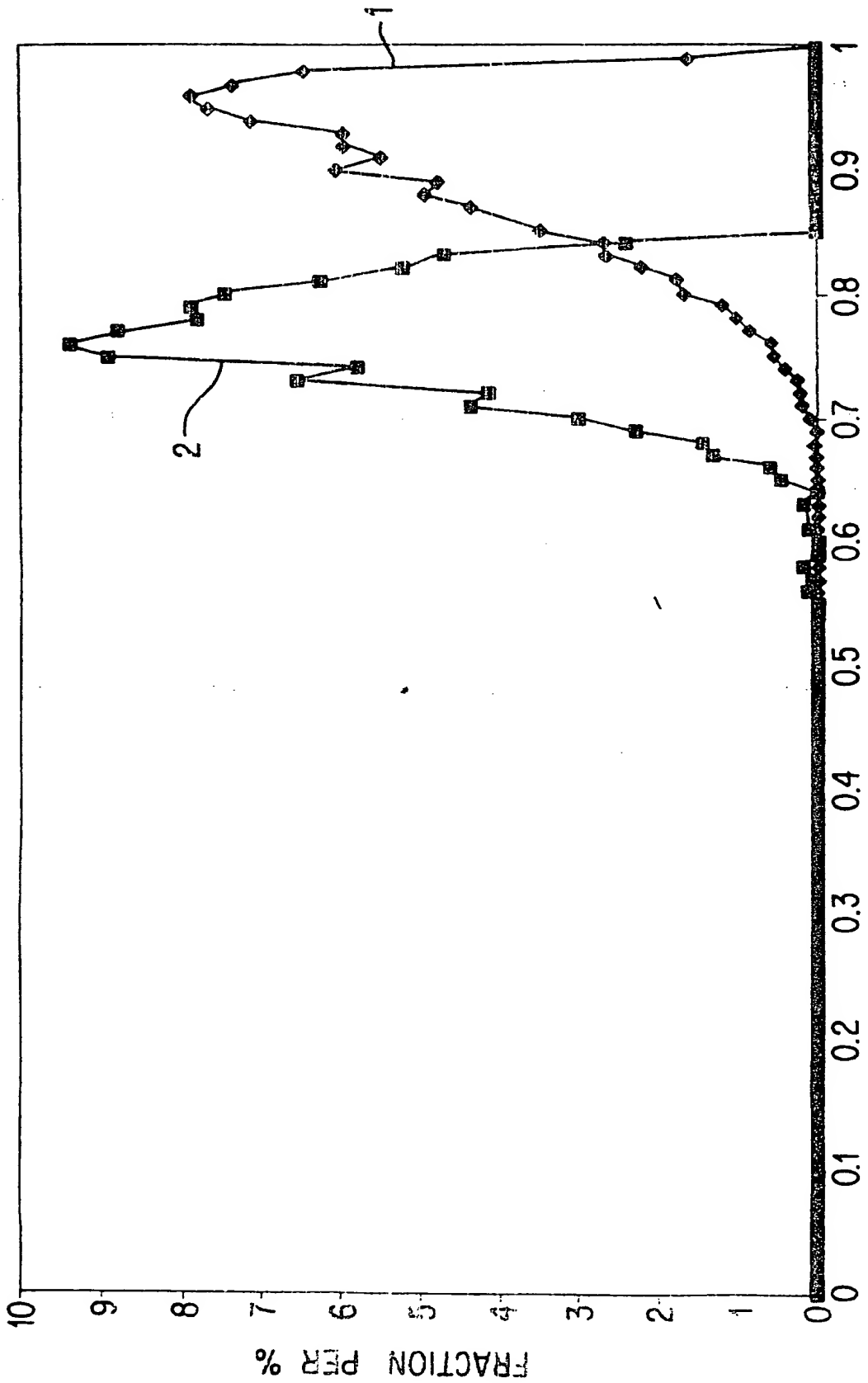

1
2
FRACTION PER %
0
1
2
3
4
5
6
7
8
9
10
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
TANIMOTO SIMILARITY

FIG. 14

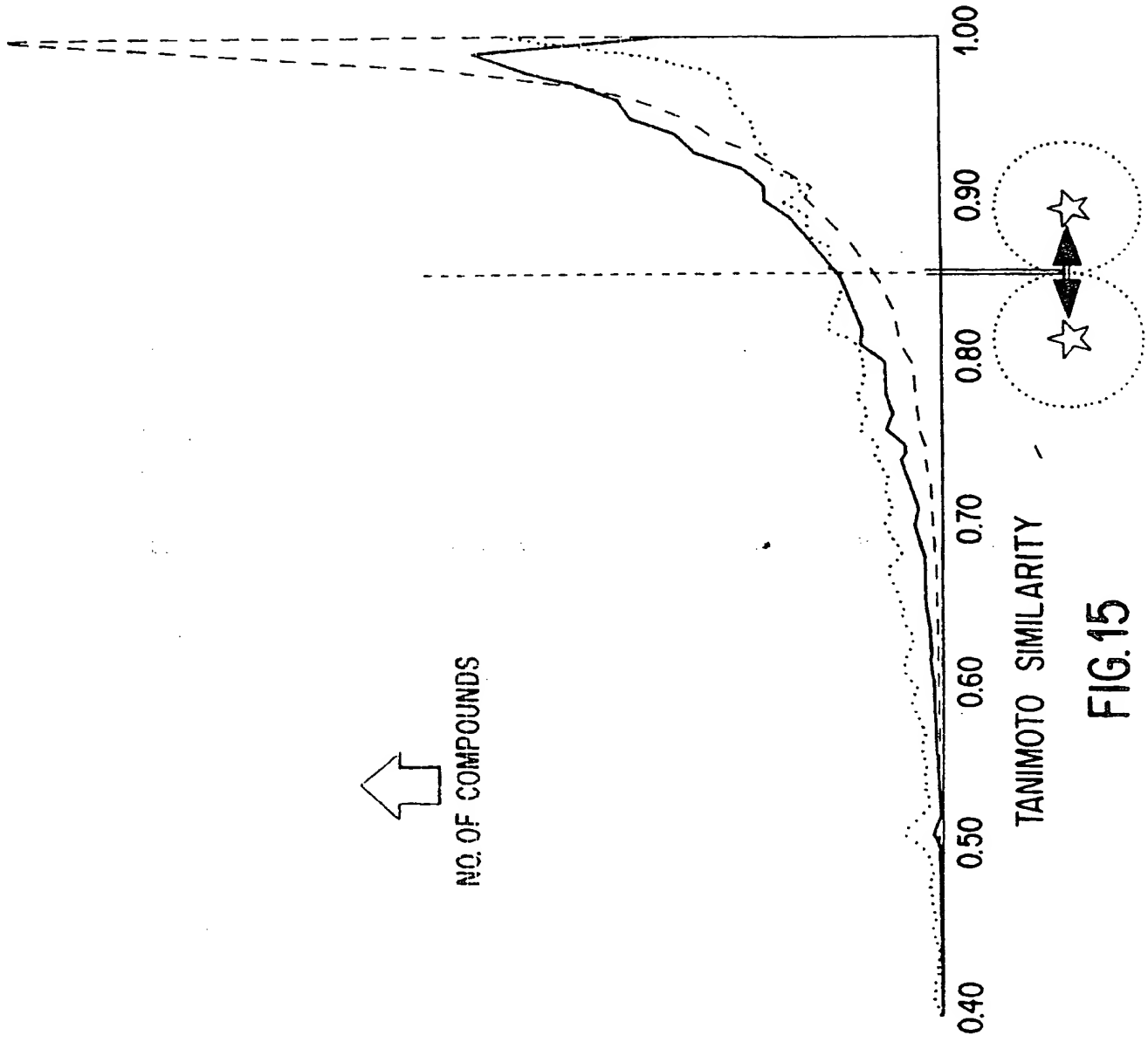

1.00
0.90
0.80
0.70
0.60
0.50
0.40
TANIMOTO SIMILARITY
NO. OF COMPOUNDS

FIG. 15

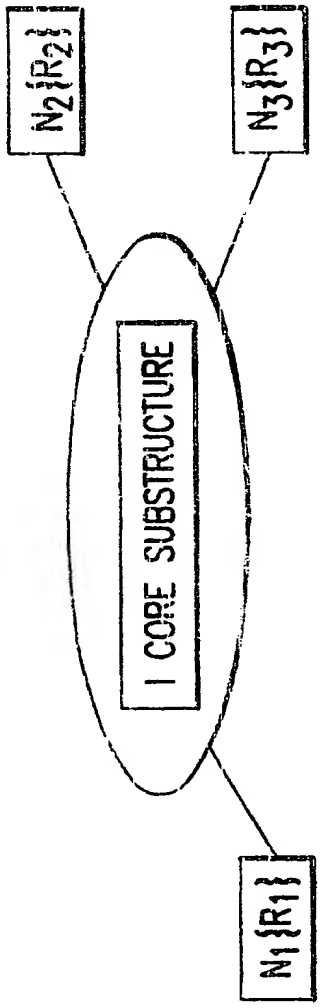


FIG. 16

FIGURE 17

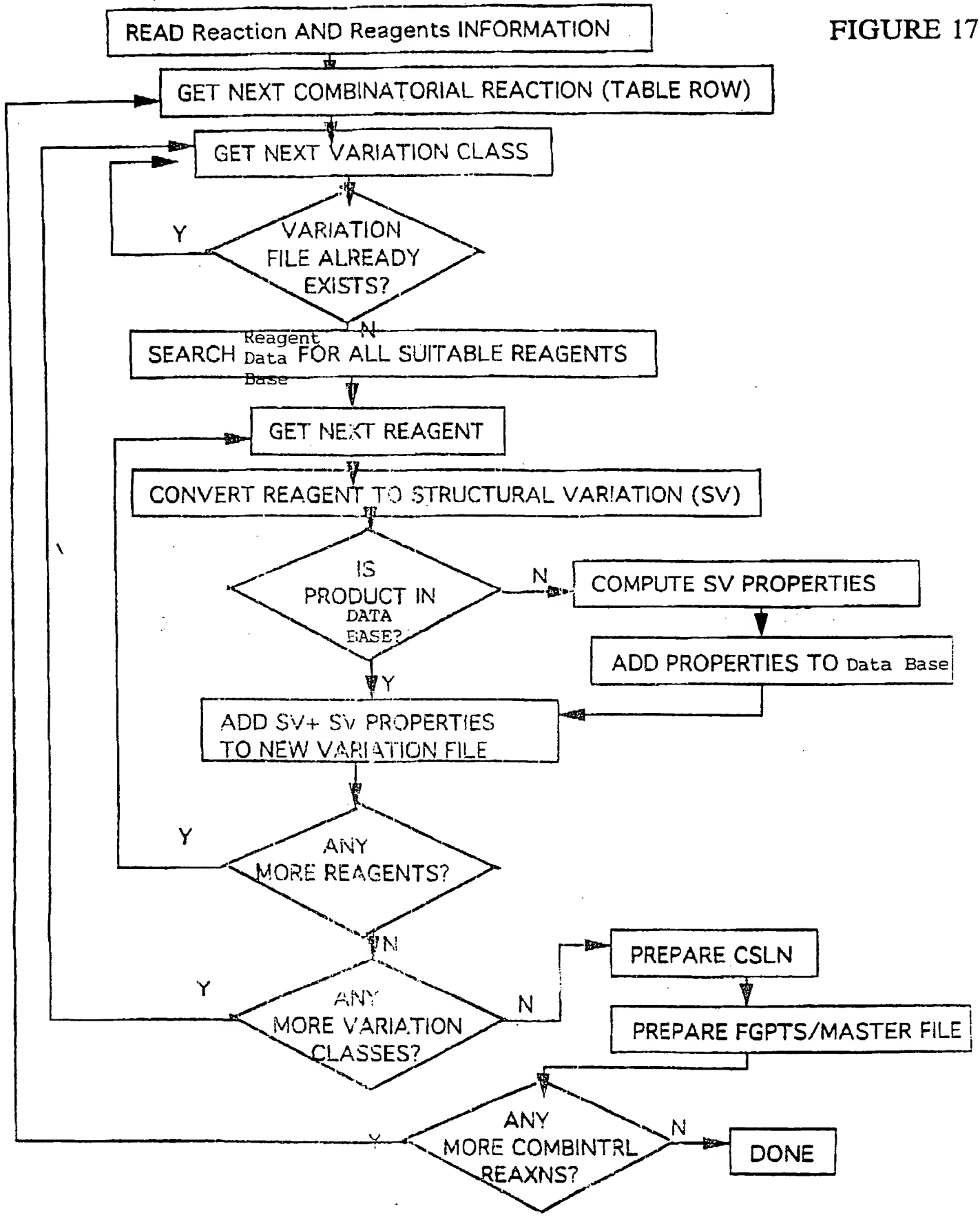

READ Reaction AND Reagents INFORMATION
GET NEXT COMBINATORIAL REACTION (TABLE ROW)
GET NEXT VARIATION CLASS
VARIATION FILE ALREADY EXISTS?
Y
N
SEARCH Reagent Data Base FOR ALL SUITABLE REAGENTS
GET NEXT REAGENT
CONVERT REAGENT TO STRUCTURAL VARIATION (SV)
IS PRODUCT IN DATA BASE?
N
COMPUTE SV PROPERTIES
ADD PROPERTIES TO Data Base
Y
ADD SV+ SV PROPERTIES TO NEW VARIATION FILE
ANY MORE REAGENTS?
Y
N
ANY MORE VARIATION CLASSES?
Y
N
PREPARE CSLN
PREPARE FGPTS/MASTER FILE
ANY MORE COMBINTRL REAXNS?
Y
N
DONE

FIGURE 18

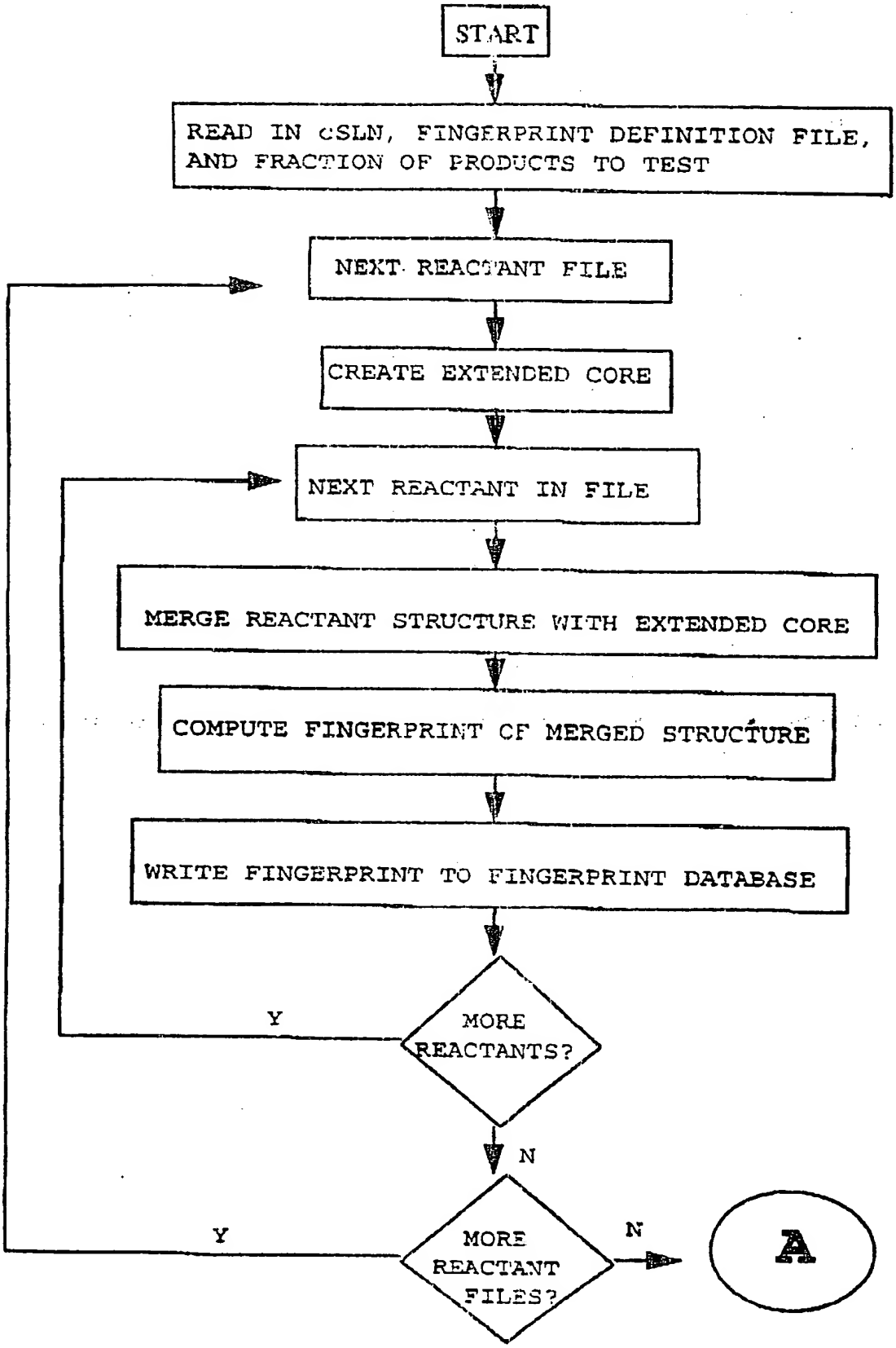

START
READ IN CSLN, FINGERPRINT DEFINITION FILE, AND FRACTION OF PRODUCTS TO TEST
NEXT REACTANT FILE
CREATE EXTENDED CORE
NEXT REACTANT IN FILE
MERGE REACTANT STRUCTURE WITH EXTENDED CORE
COMPUTE FINGERPRINT OF MERGED STRUCTURE
WRITE FINGERPRINT TO FINGERPRINT DATABASE
MORE REACTANTS?
Y
N
MORE REACTANT FILES?
Y
N
A

FIGURE 19

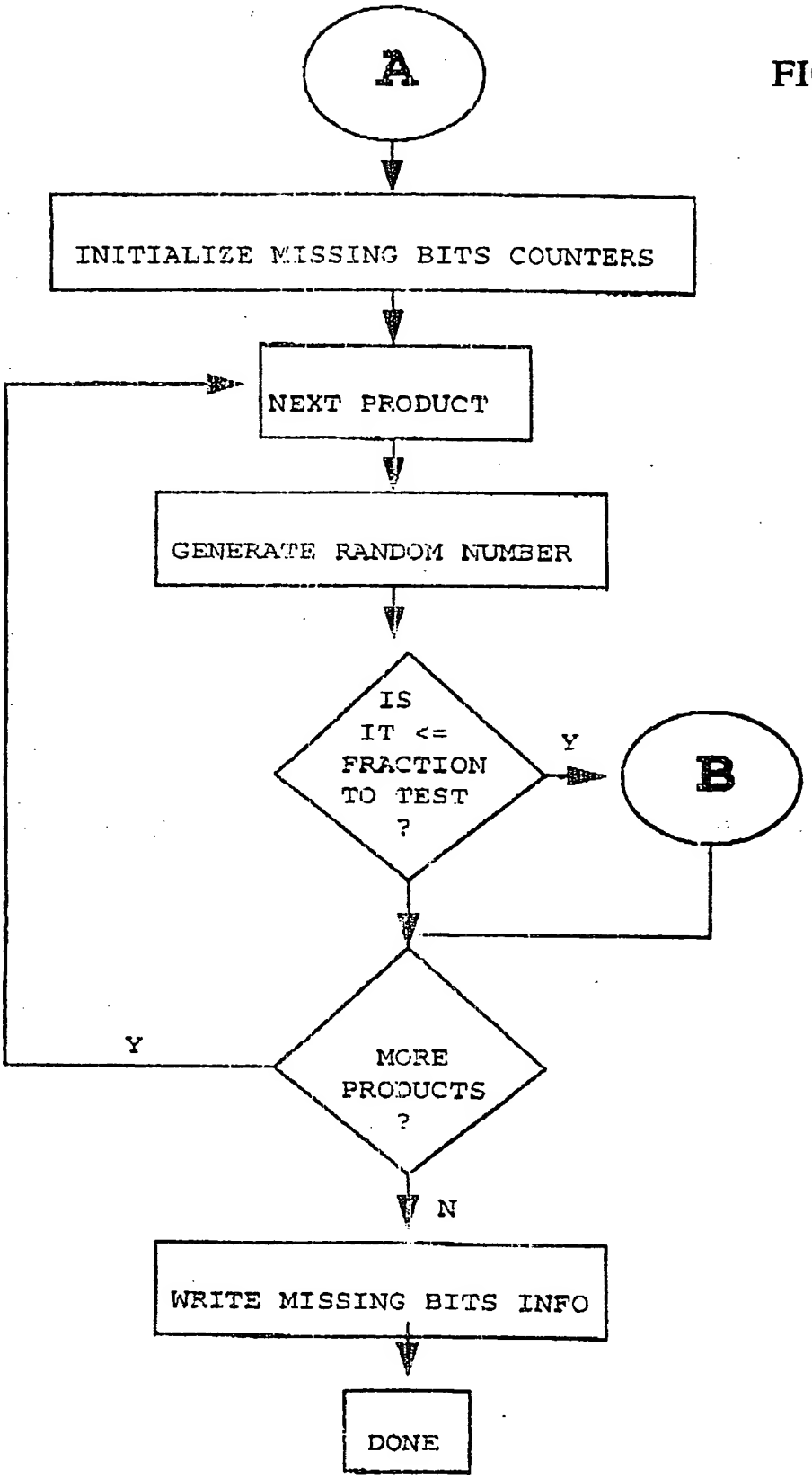

A
INITIALIZE MISSING BITS COUNTERS
NEXT PRODUCT
GENERATE RANDOM NUMBER
IS IT <= FRACTION TO TEST ?
Y
B
MORE PRODUCTS ?
Y
N
WRITE MISSING BITS INFO
DONE

FIGURE 20

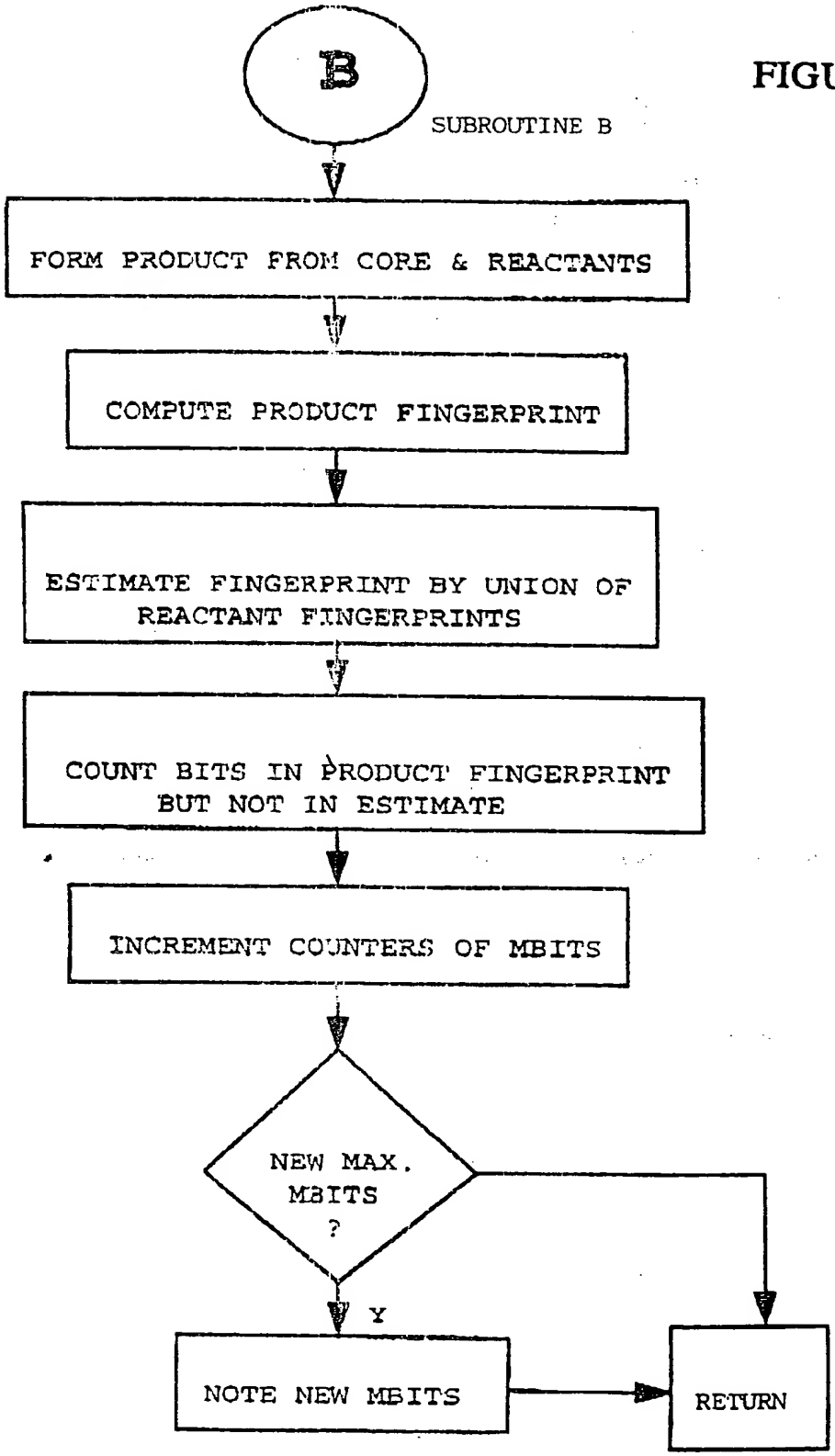

B
SUBROUTINE B
FORM PRODUCT FROM CORE & REACTANTS
COMPUTE PRODUCT FINGERPRINT
ESTIMATE FINGERPRINT BY UNION OF REACTANT FINGERPRINTS
COUNT BITS IN PRODUCT FINGERPRINT BUT NOT IN ESTIMATE
INCREMENT COUNTERS OF MBITS
NEW MAX. MBITS ?
Y
NOTE NEW MBITS
RETURN

FIGURE 2

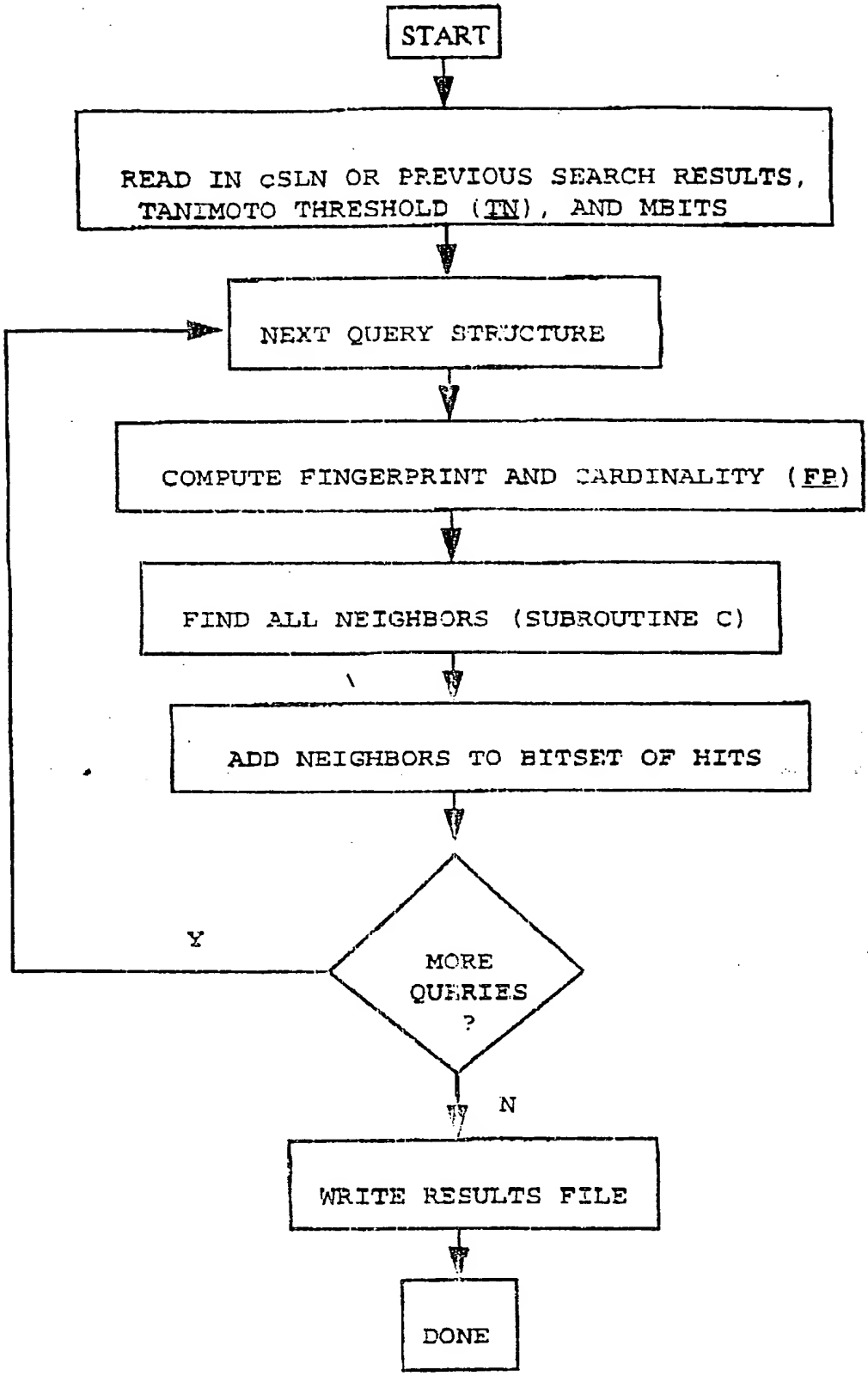

START
READ IN cSLN OR PREVIOUS SEARCH RESULTS, TANIMOTO THRESHOLD (TN), AND MBITS
NEXT QUERY STRUCTURE
COMPUTE FINGERPRINT AND CARDINALITY (FP)
FIND ALL NEIGHBORS (SUBROUTINE C)
ADD NEIGHBORS TO BITSET OF HITS
MORE QUERIES ?
Y
N
WRITE RESULTS FILE
DONE

FIGURE 22

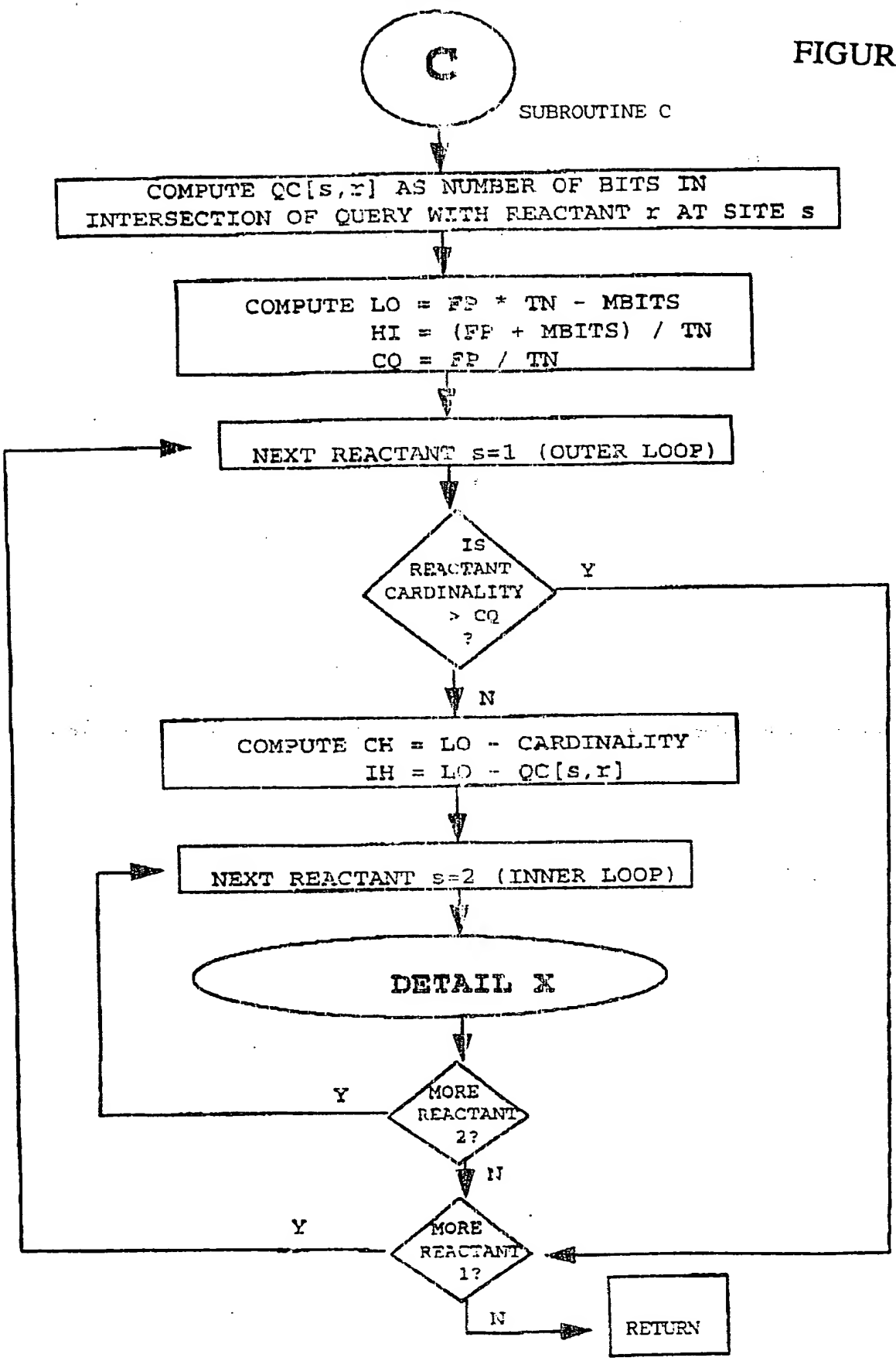

C
SUBROUTINE C
COMPUTE QC[s,r] AS NUMBER OF BITS IN INTERSECTION OF QUERY WITH REACTANT r AT SITE s
COMPUTE LO = FP * TN - MBITS
HI = (FP + MBITS) / TN
CQ = FP / TN
NEXT REACTANT s=1 (OUTER LOOP)
IS REACTANT CARDINALITY > CQ ?
Y
N
COMPUTE CH = LO - CARDINALITY
IH = LO - QC[s,r]
NEXT REACTANT s=2 (INNER LOOP)
DETAIL X
MORE REACTANT 2?
Y
N
MORE REACTANT 1?
Y
N
RETURN

FIGURE 23

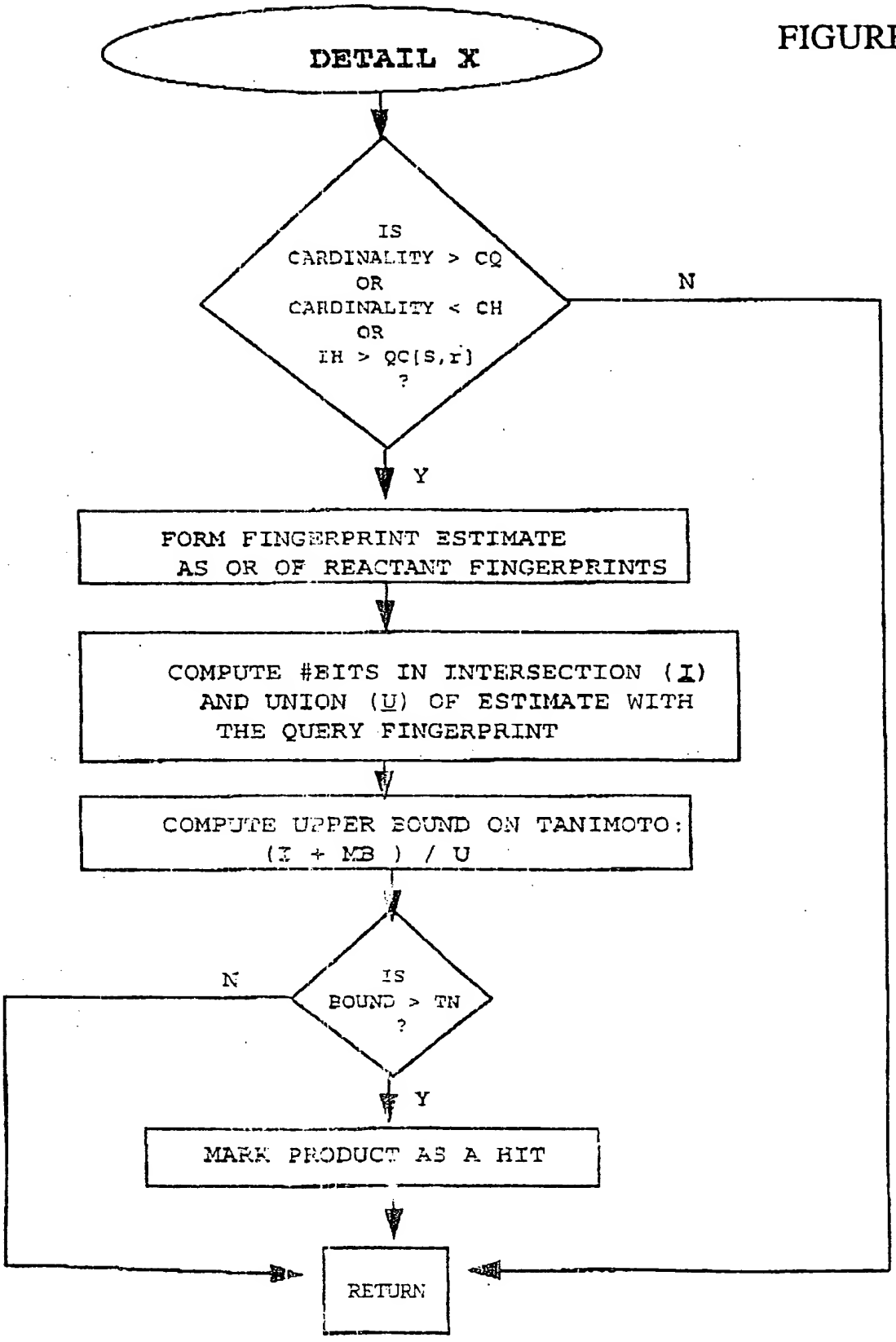

DETAIL X
IS CARDINALITY > CQ OR CARDINALITY < CH OR IH > QC[S,r] ?
N
Y
FORM FINGERPRINT ESTIMATE AS OR OF REACTANT FINGERPRINTS
COMPUTE #BITS IN INTERSECTION (I) AND UNION (U) OF ESTIMATE WITH THE QUERY FINGERPRINT
COMPUTE UPPER BOUND ON TANIMOTO: (I + MB) / U
IS BOUND > TN ?
N
Y
MARK PRODUCT AS A HIT
RETURN

FIGURE 24

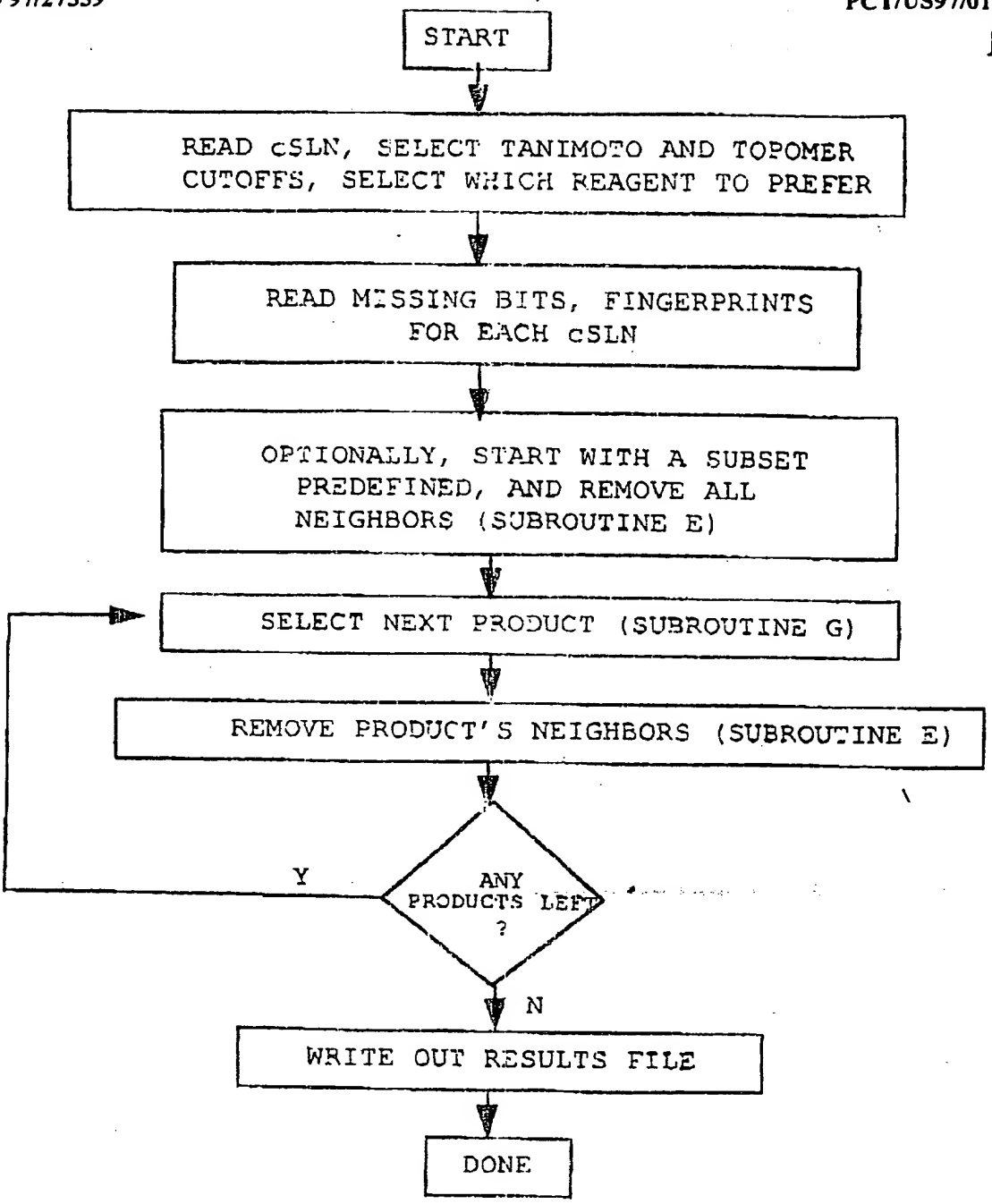

START
READ cSLN, SELECT TANIMOTO AND TOPOMER CUTOFFS, SELECT WHICH REAGENT TO PREFER
READ MISSING BITS, FINGERPRINTS FOR EACH cSLN
OPTIONALLY, START WITH A SUBSET PREDEFINED, AND REMOVE ALL NEIGHBORS (SUBROUTINE E)
SELECT NEXT PRODUCT (SUBROUTINE G)
REMOVE PRODUCT'S NEIGHBORS (SUBROUTINE E)
ANY PRODUCTS LEFT ?
Y
N
WRITE OUT RESULTS FILE
DONE

FIGURE 25

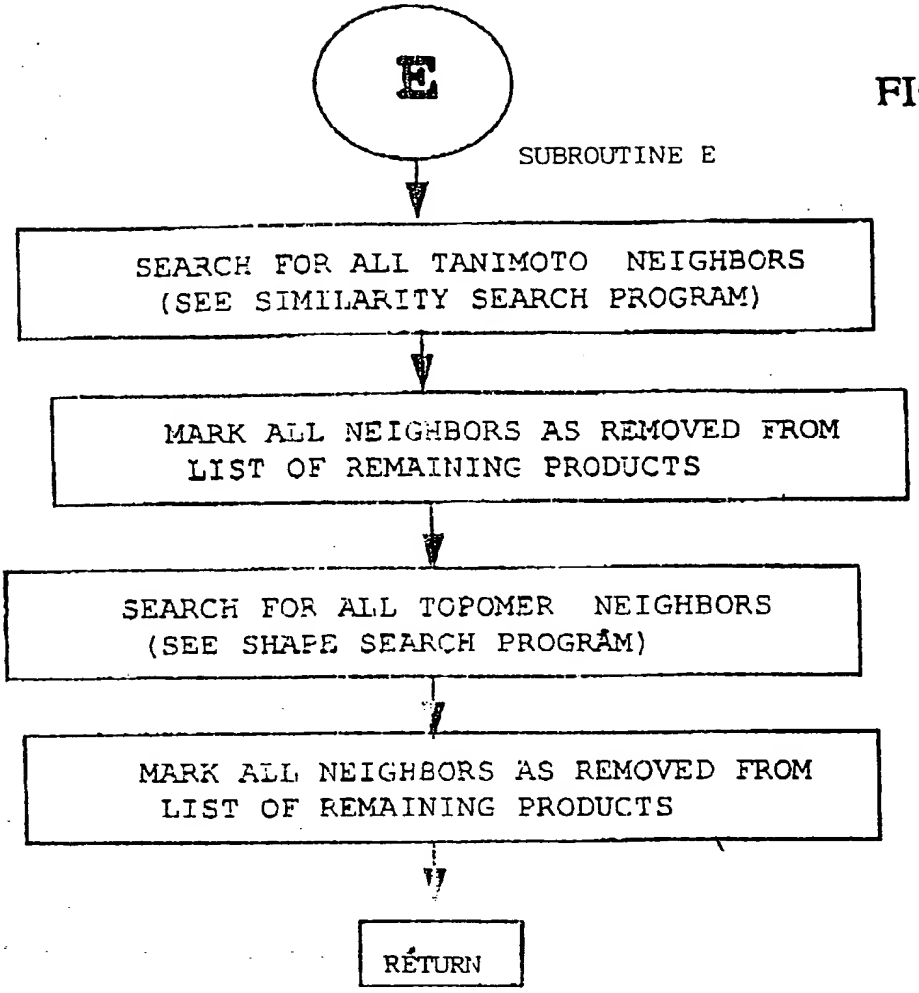

E
SUBROUTINE E
SEARCH FOR ALL TANIMOTO NEIGHBORS
(SEE SIMILARITY SEARCH PROGRAM)
MARK ALL NEIGHBORS AS REMOVED FROM
LIST OF REMAINING PRODUCTS
SEARCH FOR ALL TOPOMER NEIGHBORS
(SEE SHAPE SEARCH PROGRAM)
MARK ALL NEIGHBORS AS REMOVED FROM
LIST OF REMAINING PRODUCTS
RETURN

FIGURE 26

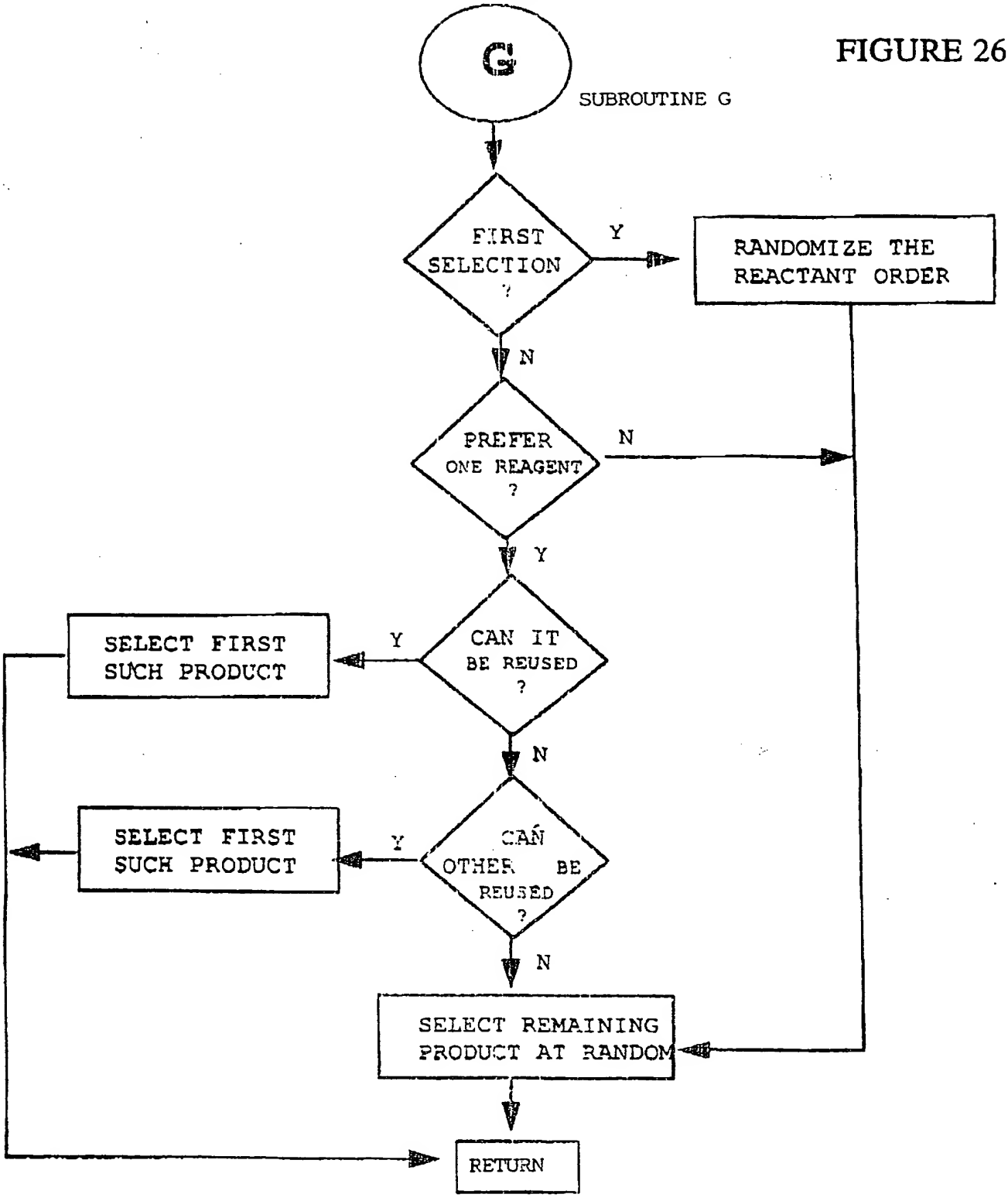

G
SUBROUTINE G
FIRST SELECTION ?
Y
RANDOMIZE THE REACTANT ORDER
N
PREFER ONE REAGENT ?
N
Y
CAN IT BE REUSED ?
Y
SELECT FIRST SUCH PRODUCT
N
CAN OTHER BE REUSED ?
Y
SELECT FIRST SUCH PRODUCT
N
SELECT REMAINING PRODUCT AT RANDOM
RETURN

FIGURE 27

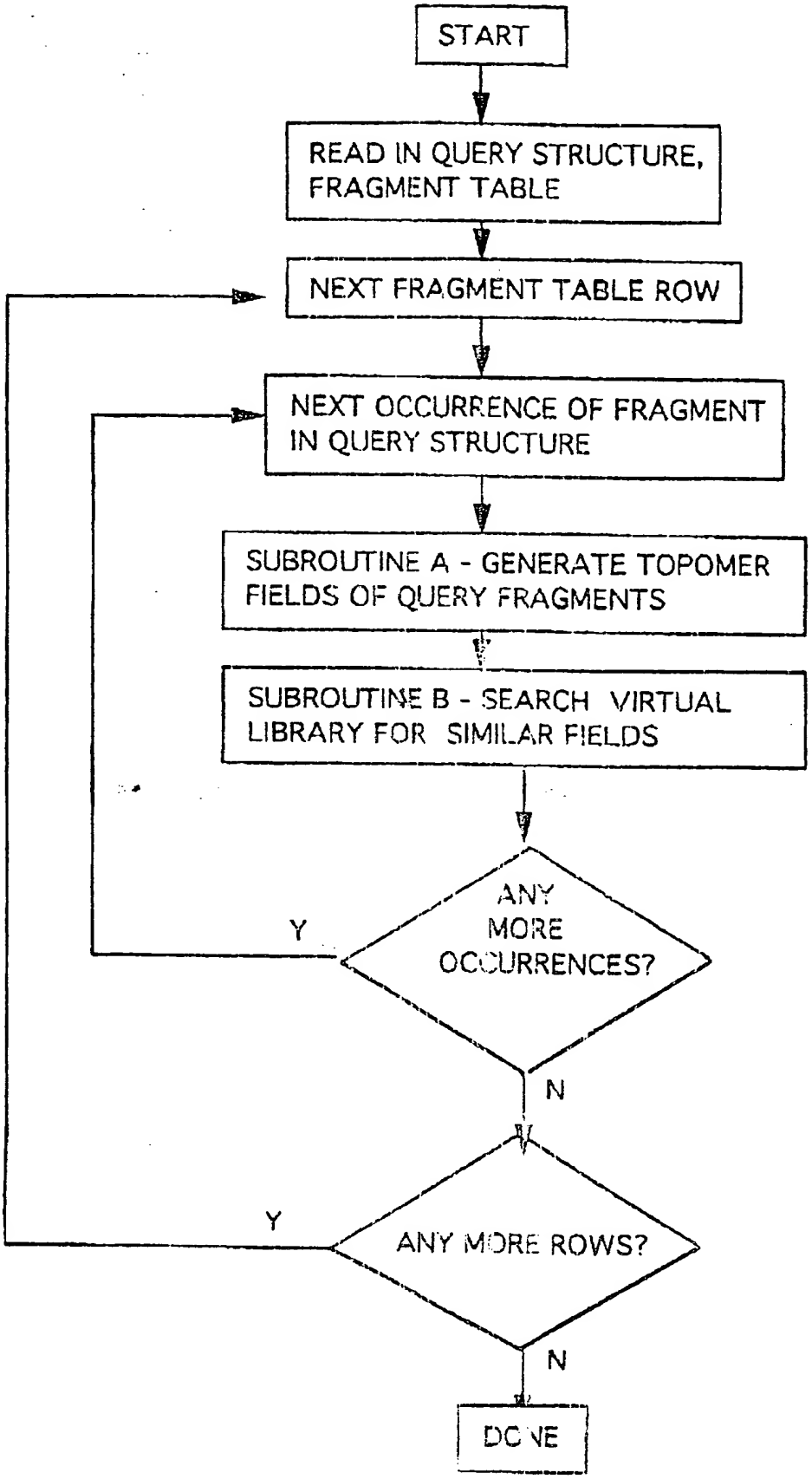
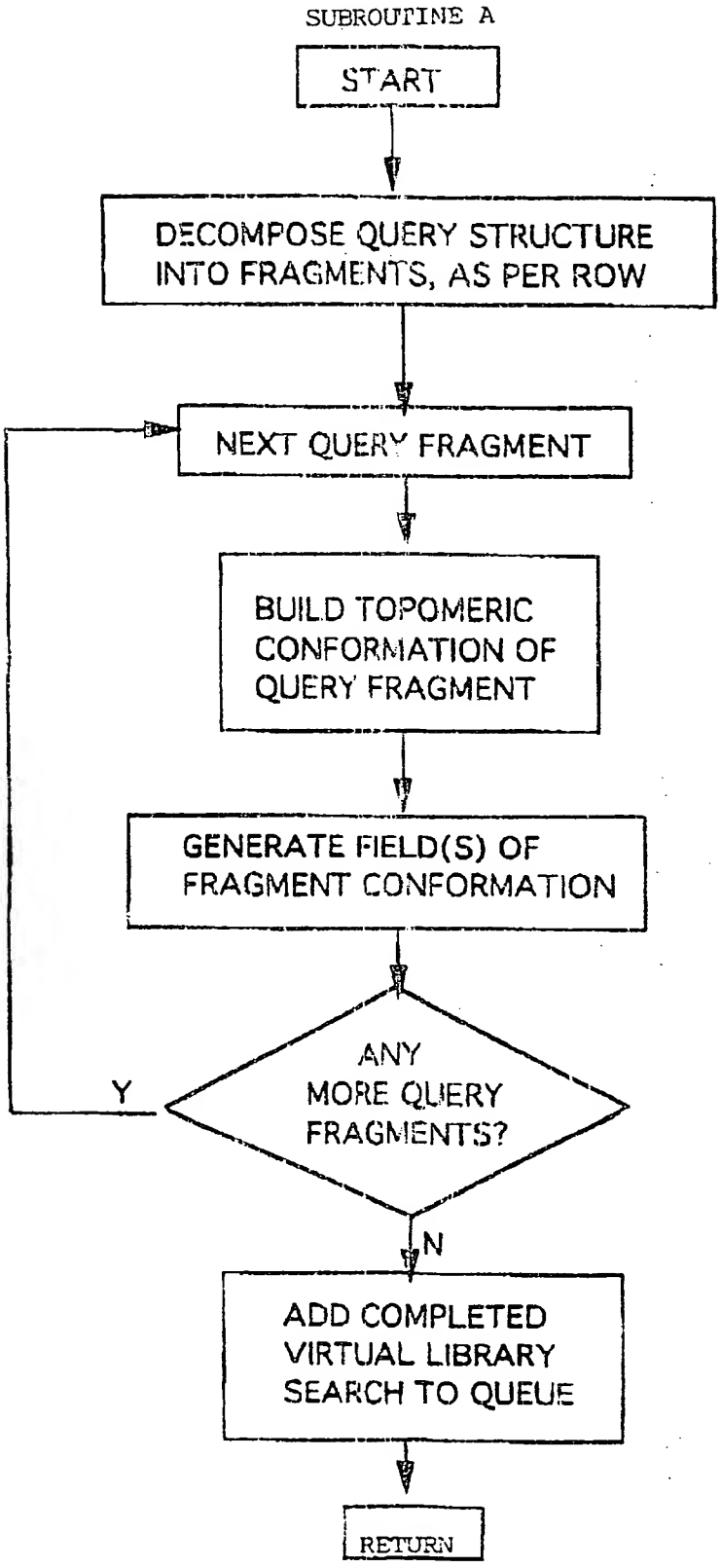
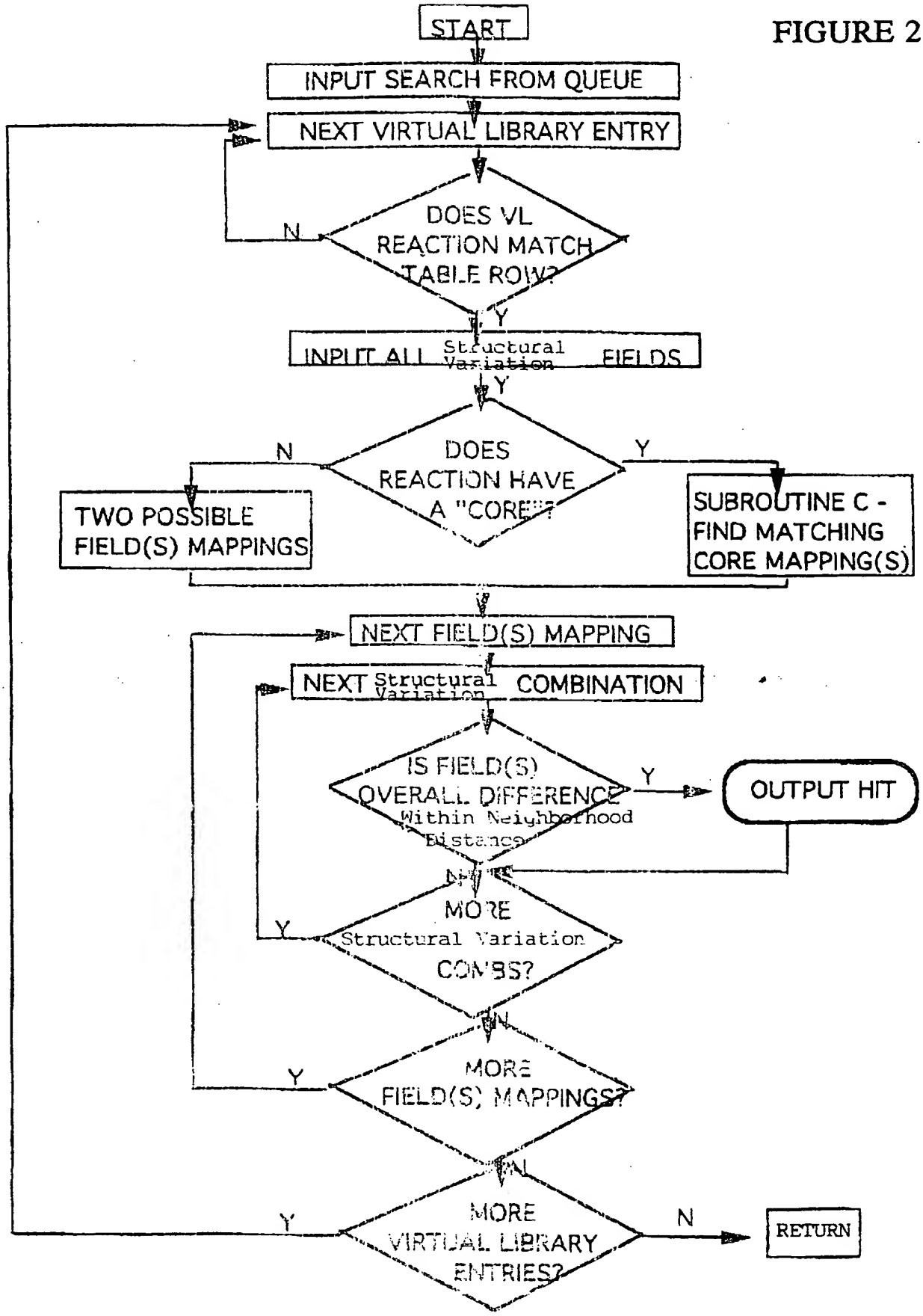

START
READ IN QUERY STRUCTURE, FRAGMENT TABLE
NEXT FRAGMENT TABLE ROW
NEXT OCCURRENCE OF FRAGMENT IN QUERY STRUCTURE
SUBROUTINE A - GENERATE TOPOMER FIELDS OF QUERY FRAGMENTS
SUBROUTINE B - SEARCH VIRTUAL LIBRARY FOR SIMILAR FIELDS
ANY MORE OCCURRENCES?
Y
N
ANY MORE ROWS?
Y
N
DONE

FIGURE 28


SUBROUTINE A
START
DECOMPOSE QUERY STRUCTURE INTO FRAGMENTS, AS PER ROW
NEXT QUERY FRAGMENT
BUILD TOPOMERIC CONFORMATION OF QUERY FRAGMENT
GENERATE FIELD(S) OF FRAGMENT CONFORMATION
ANY MORE QUERY FRAGMENTS?
Y
N
ADD COMPLETED VIRTUAL LIBRARY SEARCH TO QUEUE
RETURN

SUBROUTINE B

FIGURE 29



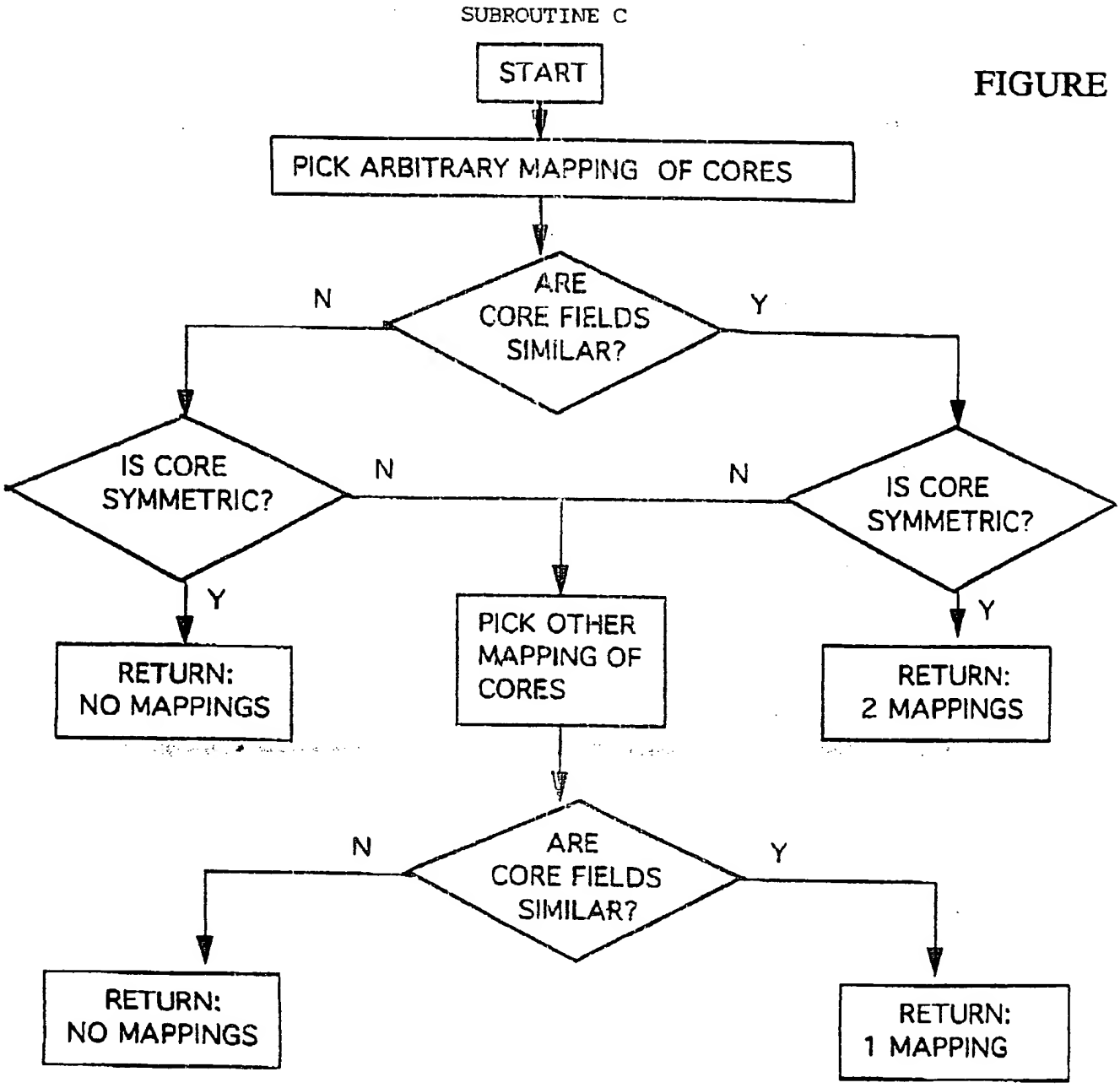
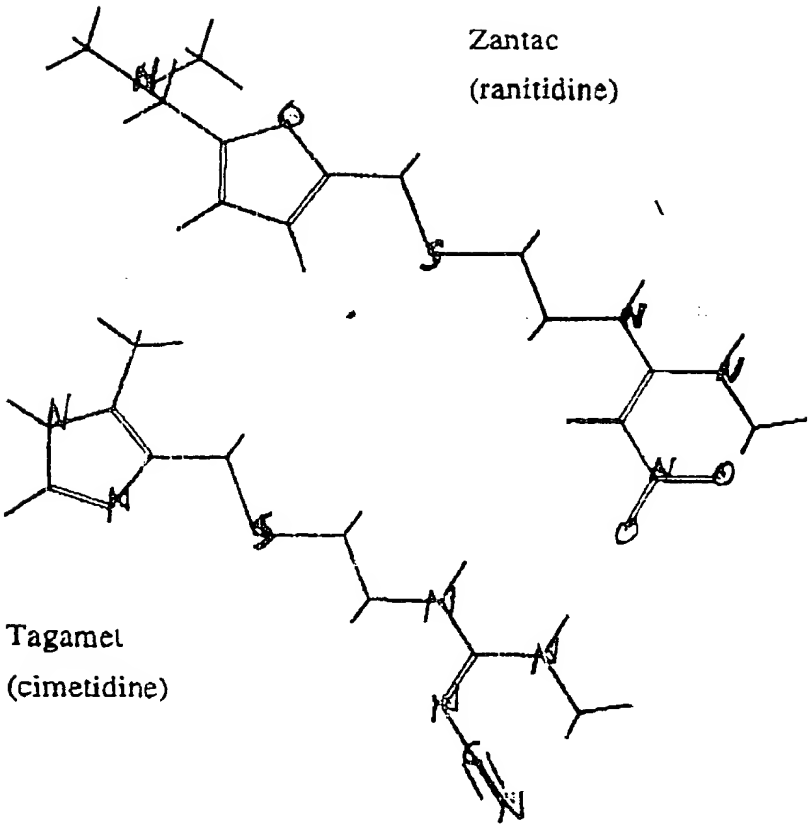

SUBROUTINE C
START
PICK ARBITRARY MAPPING OF CORES
ARE CORE FIELDS SIMILAR?
N
Y
IS CORE SYMMETRIC?
N
Y
RETURN: NO MAPPINGS
IS CORE SYMMETRIC?
N
Y
RETURN: 2 MAPPINGS
PICK OTHER MAPPING OF CORES
ARE CORE FIELDS SIMILAR?
N
Y
RETURN: NO MAPPINGS
RETURN: 1 MAPPING

FIGURE 30

FIGURE 31



INTERNATIONAL SEARCH REPORT

International application No.
PCT/US97/01491

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G06F 19/00
US CL :364/496

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 364/496-499; 395/601,616

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,307,287 A (CRAMER, III ET AL) 26 April 1994, see abstract.	52-54
X	US 5,025,388 A (CRAMER, III ET AL) 18 June 1991, see abstract.	52-54
A	US 5,345,516 A (BOYER ET AL) 09 September 1994, see entire document.	1-94
A	US 5,270,170 A (SCHATZ ET AL) 14 December 1993, see entire document.	1-94

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
21 APRIL 1997	28 MAY 1997
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer EMANUEL T. VOELTZ Telephone No. (703) 305-9714

Form PCT/ISA/210 (second sheet)(July 1992)*